STATISTICAL MODELING AND EXPERIMENTAL DESIGN WITH CONTRIBUTIONS IN ENVIRONMENT, HEALTH CARE, AND E-COMMERCE

A Dissertation
Presented to
The Academic Faculty

By

Yuanshuo(David) Zhao

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of School of Industrial and System Engineering

Georgia Institute of Technology

May 2019

STATISTICAL MODELING AND EXPERIMENTAL DESIGN WITH CONTRIBUTIONS IN ENVIRONMENT, HEALTH CARE, AND E-COMMERCE

Approved by:

Dr. C. F. Jeff Wu
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Benjamin Haaland
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Roshan Vengazhiyil
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yajun Mei
School of Industrial and Systems Engineering
Georgia Institute of Technology

Dr. Yu Jeffrey Hu
Scheller College of Business
Georgia Institute of Technology

Date Approved: March 1st, 2018

To my Mom Huiping Ma and my Dad Yongliang Zhao.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Professor C.F.Jeff Wu, for his guidance and immense support. Professor Jeff Wu is not only a great statistician, but also a great scholar and human being. I certainly would have not been here without your devoted encouragement and patience. You have not only been my academic advisors, but also incredible mentors and role models to me. It has been a privilege to work alongside you.

Second, I would like to thank my co-advisor Professor Benjamin Haaland. As one of the best junior statistician, Dr.Haaland has been completely supportive of my work and aspirations. Even after leaving Atlanta, Dr.Haaland has tried everything to support my work and never runs out of patience even when I was a bit slow. Thank you for being a mentor and always greeting me with a smile and handshake.

I also want to thank Professors Roshan Vengazhiyil, Yajun Mei and Yu Jeffrey Hu for generously offering their time to serve as committee members. Your support and kindness throughout my doctoral studies means a lot to me. Special thanks to my officemate and collaborator Simon Mak, it has been a pleasure to work with you.

I would like to thank the many wonderful faculty and staff members in ISyE for giving me the continued opportunity and support to succeed at all phases of my doctoral program. In particular, I want to thank Professor Alan Erera for providing guidance throughout my Ph.D. study and Professor Chen Zhou for academic and career advise since I was a undergraduate student at ISyE. Special thanks to Amanda Ford for making my life at ISyE simple and enjoyable.

I would like to thank the many friends and colleagues that I have made during my five years as a Ph.D. student at Georgia Tech. In particular, I want to thank Chih-li Sung, Simon Mak, Zhehui Chen, Li-Hsiang Lin, and Wenjia Wang for being a great officemates and friends. There are countless other wonderful people in ISyE who made this experience worthwhile; I will be forever grateful for our time together.

Last, but definitely not least, my appreciation and gratitude goes to my parents and my girlfriend Jing Qin for their continuous love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Design of experiment and statistical modeling have played an increasingly important role in science and business and received enormous attention from industries and research institutes. Motivated from real-world examples, this dissertation develops new statistical methodologies in the field of experimental design and causality inferences. First two chapters of this dissertation focus on online experimental design. E-commerce companies like Linkedin and Amazon perform hundreds of experiments each day, with the goal of testing certain website functions and design in order to best serve customers and maximize profits. New experiment design and testing scheme based on multi-armed bandit and conditional main-effect have been developed to let companies run experiment more efficiently. In chapter three, we develop a new statistical model based on combining information from physical experiment and computer experiment. The new method has been applied to model the Solar Irradiance data in the U.S. that were provided by IBM. Chapter four extends the linear G-formula method in the field of causality inference to non-linear set-up to study the causality relationship between physical activity level and health outcomes.

In e-commerce companies, a key step for revenue optimization is designing a website which maximizes conversion rates. This is achieved by first running many conversion experiments on different website settings (i.e., with different combinations of design factors), then using this data to pick an optimal website setting. In real-world scenarios, there are oftentimes many factors of interest, resulting in a large website design space. For such problems, only a small fraction of websites can be run in each experiment round due to budget constraints. This poses a problem for traditional multi-armed bandit methods, which

typically assume all website settings (arms) are tested in each experiment round. To address this so-called ”arm budget constraint”, in chapter 1, we propose a new method called Active Arm SElection using Thompson Sampling (AASETS), which performs active arm selection and traffic allocation in an online setting, under a fixed budget of arms in each experiment round. The key novelty of AASETS is the use of a low-order interaction model to learn dependencies between arms on the factorial design space. This model allows an experimenter to (i) adaptively add good arms and remove bad arms from experimentation, and (ii) leverage conversion data over all arms for effective traffic allocation. We show that AASETS outperforms several industry benchmark methods by a large margin under arm budget constraints, both in simulated examples and a real-world problem.

Chpter 2 proposed a new statistical testing method based on conditional main-effect for conversion rate optimization. Conversion rate optimization has become more important because of the rapid growth of e-commerce revenue. Traditional conversion rate optimization, including AB testing and multivariate testing, tends to isolate factors and treat them the same regardless of their positions in the web system. In this chapter, we will discuss a new framework, called Conditional main-effect based funnel testing, where factors effects and level settings are analyzed and optimized based on their position on the webpage. We called the new approach CFO: Conditional effect based Funnel testing for conversion rate optimization. The new approach has better interpretability of the factorial effect and achieves better result in conversion rate optimization.

The Gaussian process is a standard tool for building emulators for computer experiments. However, due to its lack of ability to model large-scale and non-stationary data, Gaussian process is greatly limited in practice. In chapter 3, We provide a new approach to approxi-

mate emulation of large computer experiments. By taking advantage of the learning ability and strong tolerance to input noise of radial basis function, we derive a sequential learning scheme that dynamically optimizes the basis function's location, scale, and coefficient. L-1 penalty is utilized to ensure our emulator's simplicity. We applied our method to study solar irradiance computer model and physical measurements data. We demonstrate that the proposed model enjoy marked advantage over existing emulation tools in both emulation accuracy and data capability in terms of non-nationality and sample size. The final predictor based on combining physical measurement data and computer experiment data is used to forecast the solar irradiance level in the U.S.

TRIPPA (trial of economic incentives to promote physical activity) was a four-arm, 6 month randomized controlled trial with a 6-month post-intervention follow-up period, conducted in 13 organizations spanning industries and sectors of government, to investigate the effects of an activity tracker, with or without cash or charitable incentives, on physical activity and health outcomes among full-time workers in Singapore. In chapter 4, we conduct a follow-up study of TRIPPA to assess the causal effects of physical activity levels on health outcomes, including systolic blood pressure (SBP), BMI, VO2MAX and quality-of-life. We extended the original g-formula framework that deals with time-varying confounding to include non-linear models, which allows us to use statistical models that are more robust compared to linear models.

CHAPTER 1

ACTIVE ARM SELECTION USING THOMPSON SAMPLING (AASETS): A MULTI-ARMED BANDIT METHOD UNDER ARM BUDGET CONSTRAINTS

1.1 Abstract

In e-commerce companies, such as Amazon and LinkedIn, a key step for revenue optimization is designing a website which maximizes conversion rates. This is achieved by first running many conversion experiments on different website settings (i.e., with different combinations of design factors), then using this data to pick an optimal website setting. In real-world scenarios, there are oftentimes many factors of interest, resulting in a large website design space. For such problems, only a small fraction of websites can be run in each experiment round due to budget constraints. This poses a problem for traditional multi-armed bandit methods, which typically assume all website settings (arms) are tested in each experiment round. To address this so-called "arm budget constraint", we propose a new method called **A**ctive **A**rm **SE**lection using **T**hompson **S**ampling (AASETS), which performs active arm selection and traffic allocation in an online setting, under a fixed budget of arms in each experiment round. The key novelty of AASETS is the use of a low-order interaction model to learn dependencies between arms on the factorial design space. This model allows an experimenter to (i) adaptively add good arms and remove bad arms from experimentation, and (ii) leverage conversion data over all arms for effective traffic allocation. We show that AASETS outperforms several industry benchmark methods by a large

margin under arm budget constraints, both in simulated examples and a real-world problem.

1.2 Introduction

Conversion rate optimization – the system for maximizing the conversion percentage of website visitors to customers – plays a central role in e-commerce. Companies such as Amazon or Linkedin perform hundreds of experiments each day, with the goal of testing certain website functions and designs in order to best serve customers and maximize profits. Currently, A/B testing and multivariate testing is the default for conversion rate optimization. In such methods, the traffic is evenly allocated, experiments are run for a certain time period, and then the ”best” setting is picked for implementation [1]. In statistical terms, the above experiment framework consists of first a pure *exploration* period, where an experimenter randomly assigns equal number of users to different website configurations. It then moves into the period of *exploitation*, where an experimenter sends all of their traffic into the most successful configurations. There are two problems with the above framework. First, the transition between *exploration* and *exploitation* is discrete. After exploration, it is unlikely that the experimenter can conclude with certainty that a particular version is the optimal one; such uncertainty may result in large losses during exploitation. Second, during exploration, resources may be wasted by allocating traffic to suboptimal choices. A method which provides a “soft” transition between exploration and exploitation (by quickly discarding inferior configurations) is therefore desired, since it can save website owners a considerable amount of experiment costs.

The above problem can be formulated as a stochastic multi-armed bandit (MAB) problem [2]. The term ’multi-armed bandit’ comes from the world of gambling. For the motivating

website design problem, imagine the different website versions as a row of slot machines, each with its own probability of producing a reward (conversion). The goal of the player (i.e., experimenter) is to accumulate as much reward as possible during the duration of the playing time. The player can choose to play the slot machine which gives the highest reward so far (i.e., allocate all of its traffic to the best website at the time (exploitation). However, the early superiority of this slot machine can be due to luck, and it may be beneficial to play other machines in hopes of getting better rewards (exploration). An early seminal paper in the MAB literature is [3], who proposed arm selection polices which enjoy the fastest convergence rate, in the case where reward distributions are in the one-parameter exponential family. [4] later offers a simplification of this policy, and proved a key convergence result for the case of normal populations with known variances. Recent work has focused on finding approximate solutions to these optimal MAB policies, the popular method being Thompson sampling [5], [6]. We will introduce Thompson sampling in greater detail later in the paper. Other notable developments of the MAB set-up includes contextual multi-armed bandits [7] and adversarial bandits [8].

However, a key assumption in the MAB literature is that an experimenter can play all the arms in an experimental period. Such an assumption may not be true in real-world e-commerce environments. For example, suppose the arms represent website configurations with different combinations of font family, font size, image location, and background color. If each of these characteristics has five choices (levels), then there are $625 (= 5^4)$ possible configurations to be tested. In practice, a company usually cannot test more than 30 website versions at the same time due to maintenance costs. Given this so-called "arm budget constraint", an experimenter would therefore like to adaptively choose which arms to play

in the current time period (we call these arms the *active set*). To our knowledge, the existing MAB literature considers mainly the "design" of the allocation scheme (i.e., how much traffic should be assigned to each arm), and largely ignores the "design" problem of active arms. However, given that we cannot feasibly experiment on all website configurations, the latter active arm design problem then plays a key role in maximizing information for conversion optimization.

To address this, we propose a new method called **A**ctive **A**rm **SE**lection using **T**hompson **S**ampling (AASETS), which performs active arm selection and traffic allocation under arm budget constraints. The main novelty of AASETS is the use of a low-order interaction model to learn about dependencies between arms. Such a model allows us to infer information (and thereby maximize reward) over a large design space, from experimental data from a limited number of observed arms. In particular, this model (when fitted to data) provides two important features of the proposed AASETS method. First, it allows an experimenter to adaptively add good arms and remove bad arms from the active set of arms, a procedure we call "switching". Second, it can leverage conversion data over all arms for effective traffic allocation. We show that AASETS outperforms existing methods in industry by a large margin, in both simulated data and real-world example under arm budget constraints.

The remainder of this paper is structured as follows. Section 2 describes and motivates the considered problem set-up, and brief reviews of the Thompson sampling allocation method. Section 3 presents the proposed AASETS framework and implementation details. Section 4 compares the performance of ASSETS to four benchmark methods commonly employed in the industry, and studies the sensitivity of our method to different modeling schemes. Section 5 then describes an application of AASETS to a real-world problem.

Section 6 concludes with a summary and future work.

1.3 Problem set-up

In the following, we briefly set-up the MAB problem under consideration, then discuss existing MAB methods, and why the proposed approach can improve upon such methods given arm budget constraints.

1.3.1 MAB under arm budget constraints

We first define some notation for MABs. Let $\mathcal{A}$ be the set of arms which can be experimented on. For a given $a \in \mathcal{A}$, let $f_a(y|\theta)$ be the distribution of the reward Y generated from arm a, where θ consists of the common parameters which govern the reward distribution over all arms. Further let $\mu_a(\theta)$ be the expected reward from distribution $f_a(y|\theta)$. Since the primary focus of this work is in tackling the motivating problem of website conversion, we assume that the reward random variable Y is binary, i.e. $Y \in \{0, 1\}$. There, a reward of 1 represents successful conversion, whereas a reward of 0 represents no conversion. Our framework can be extended in a straight-forward manner for continuous reward distributions.

In the case of binary rewards, it is clear that if the true distribution $f_a(y|\theta)$ is known, then the optimal strategy would be to always pick the arm with the largest expected reward $\mu_a(\theta)$. However, the true reward distributions are never known in practice, and it would be useful to measure the regret of making a suboptimal decision. Suppose the experiment is conducted over a timeframe of T time periods, with $n_{a,t}$ the number of traffic runs allocated to arm a at time t. Following the standard MAB literature, the *regret* L_t at time t is defined

as:

$$L_t = \sum_{a \in \mathcal{A}} n_{at}(\mu^*(\theta) - \mu_a(\theta)), \tag{1.1}$$

where $\mu^*(\theta) = \max_{a \in \mathcal{A}} \mu_a(\theta)$ is the expected reward of the optimal arm. In words, L_t quantifies the expected loss of reward resulting from a suboptimal selection of arms. The expected *reward* at time t can similarly be defined as:

$$R_t = \sum_{a \in \mathcal{A}} n_{a,t}\mu_a(\theta). \tag{1.2}$$

From this, one can then define the *cumulative regret* L over T time periods as:

$$L = \sum_{t=1}^{T} L_t. \tag{1.3}$$

For the website conversion problem, L quantifies the expected number of lost conversions over the entire experimental period. In a similar fashion, the *cumulative reward* R over T time periods as:

$$R = \sum_{t=1}^{T} R_t. \tag{1.4}$$

In this paper, we focus on the goal of minimizing cumulative regret L, or equivalently, maximizing cumulative reward R.

The experimental framework described in Section 1 for website conversion has two key distinctions from the standard MAB framework in the literature. First, within any given time period t, an experimenter can only afford to test a subset of K arms from the total $N = \#\mathcal{A}$ arms due to budget constraints, with $K \ll N$. In website optimization, this constraint

arises from a variety of practical concerns: website design costs may be expensive, and a minimum amount of traffic may be desired in each tested design to (i) ensure reliable results, and (ii) identify potential bugs (root causes for unusual traffic behavior). Second, because of this so-called *arm budget constraint*, we will allow a total of S "switches", where an experimenter can remove old arms which are likely suboptimal, and add new arms which are promising. In particular, we assume that there are exactly T time periods between switches, which corresponds to a total experimental timeframe of ST time periods.

We illustrate this experimental framework in the following practical example. Suppose a company can run batch conversion experiments in eight hourly time periods within a day (e.g., one batch per hour in a 9am - 5pm day). This corresponds to a choice of $T = 8$. In practice, companies usually switch the current set of websites (i.e., arms) at the end of each day, when traffic is at its lowest. Given an experimental timeframe of 14 days, this corresponds to a total number of $S = 14$ switches.

Faced with this practical arm budget constraint, we require a slight adaptation of regret for the new experimental framework. Within switch period s, let $\mathcal{A}_s \subset \mathcal{A}$ denote the subset of arms which are selected for experimentation (with $\mathcal{A}_s = K$). Furthermore, within the same switch period, let $n_{a,t,s}$ denote the number of traffic runs allocated to arm a in time period t. Note that, for arms $a \notin \mathcal{A}_s$ which are not selected in switch s, $n_{a,t,s}$ must equal 0. With this, the *cumulative regret* then becomes:

$$L = \sum_{s=1}^{S} \sum_{t=1}^{T} \sum_{a \in \mathcal{A}_s} n_{a,t,s}(\mu^*(\theta) - \mu_a(\theta)). \tag{1.5}$$

The goal of our work is to develop an active "design" strategy for maximizing (1.5). This

design strategy consists of two parts: the addition / removal of arms at each switch, and the allocation of traffic runs within each time period.

1.3.2 Existing MAB methods

As mentioned in the Introduction, most of the existing MAB literature ignores this arm budget constraint, and instead assume that traffic can be allocated to all arms in $\mathcal{A}$. As such, these works consider only the "design" problem of how to best allocate traffic runs to each arm in $\mathcal{A}$, in order to minimize regret. Below, we discuss in detail a popular allocation method called *Thompson sampling* [6], which we will employ within the proposed AASETS algorithm.

The key idea in Thompson sampling is to allocate traffic to arm a using the probability that arm a is the optimal arm. Define $w_a = Pr(\mu_a = max\{\mu_1, \mu_2, ..., \mu_N\})$ as the posterior probability that arm a is the optimal arm, under an appropriate Bayesian model. For some families of reward distributions, it is possible to compute w_a analytically, but in most cases it is only feasible to estimate w_a by simulations. In practice, one typically simulate $a \sim w_a$ by simulating a single draw of parameters θ_t from the posterior distribution $p(\theta|\text{data})$ and select the next arm a^* which maximizes $a^* = \arg\max_a \mu_a(\theta_t)$. One nice feature of Thompson sampling is that it relies solely on posterior draws of θ, which can be readily obtained for almost any reward distribution using Monte Carlo Markov Chain [9]. This computational feature and its appealing asymptotic properties are the primary reasons why our proposed algorithm is based on the Thompson sampling.

We demonstrate Thompson sampling using the following simple illustration. Suppose there are K arms. During the play, arm k produces a success $y = 1$ with probability θ_k

and failure $y = 0$ with probability $1 - \theta_k$. Each θ_k can be interpreted as the probability for conversion. In this paper, we term the Thompson sampling algorithm for binomial bandit [6] as "vanilla Thompson sampling". The algorithm for "vanilla" Thompson sampling maintains Bayesian priors on the Bernoulli mean reward. In practice, independent Beta(1,1) priors are often chosen for Bernoulli rewards, because it provides a non-informative conjugate prior. The probability distribution function of $Beta(\alpha, \beta)$, the beta distribution with parameter $\alpha > 0$, $\beta > 0$, is given by $f(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$. After observing the Bernoulli trial, if the trial is a success, then the posterior distribution becomes $Beta(\alpha+1, \beta)$; if the trial is a failure, then the posterior distribution becomes $Beta(\alpha, \beta+1)$. The complete Thompson sampling for Bernoulli Bandits is outlined in Algorithm 1 below:

Algorithm 1: Thompson sampling for Bernoulli Bandit

For each arm $i = 1, ..., N$, set $\alpha_i = 0$, $\beta_i = 0$;
foreach $t = 1, 2, ...$ **do**
  For each arm $i = 1, ..., N$, sample $\theta_i(t)$ from the posterior $Beta(\alpha_i + 1, \beta_i + 1)$;
  Play arm $i(t) := argmax_i \theta_i(t)$ and observe reward y;
  If $y = 1$, then $\alpha_i = \alpha_i + 1$, else $\beta_i = \beta_i + 1$

Thompson sampling enjoys many desirable theoretical properties [10], such as asymptotical optimality. It is also intuitively appealing, in that it naturally balances (i) exploration of the arm space $\mathcal{A}$, and (ii) exploitation of arms with high empirical conversions. At the beginning of Thompson sampling, the variance of the sampled Beta distribution is large, which encourages the exploration of different arms in $\mathcal{A}$. As Thompson sampling runs longer (and sample size increases), the sampled Beta distribution decreases in variance, which then encourages the exploitation of arms with high empirical conversion rates. In doing so, sub-optimal arms will then receive fewer and fewer traffic, which is as desired.

Besides Thompson sampling, there have been many different techniques proposed for solving the traffic allocation problem in multi-armed bandits. One such approach is the Gittins index [11]. While this approach is known to be theoretically optimal, it is rarely used in practice due to the demanding computational complexity of the algorithm. Another approach is the ϵ-greedy method [12], which strives to balance exploitation and exploration based on a tuning parameter ϵ. The upper confidence bound (UCB) strategies are also quite popular in the literature. UCB methods are based on the observation in [13] that upper confidence bounds in MABs, when used for allocation, efficiently approximate the Gittins index. A detailed review of the above methods can be found in [2].

1.4 AASETS: MAB under arm budget constraints

With this in hand, we now present a new method, called AASETS, which makes use of a fitted low-order interaction model to learn about dependencies between arms on the factorial design space $\mathcal{A}$. The dependencies fitted from this model allows us to infer information on the many unobserved arms in $\mathcal{A}$ by leveraging experimental data from the few observed arms. We first propose the low-order interaction modeling framework, then present a Bayesian specification for posterior sampling, and finally show how this posterior learning can guide both the traffic allocation and active arm selection.

1.4.1 Low-order interaction modeling

The motivation for this low-order interaction modeling framework comes from a recent work by [14], where an additive logistic model (with probit link function) is used to quantify the relationship between different design factors (inputs) and conversion rates (output). To see

why this may be effective in MAB problems with many arms, consider the following website design example. Suppose the arms are websites with different combinations of four design factors: font size, image location, button shape, and background color. If each of these factor have 5 levels, then there are $625(= 5^4)$ possible website configurations (arms) to test. With 625 arms, however, it is nearly impossible for the "vanilla" Thompson sampling discussed earlier (or indeed, most multi-armed bandit algorithms) to converge, given a reasonable experimental budget. To address this, [14] propose the following main-effect (or additive) model for conversion rates:

$$\mu_a(\theta) = \Phi(\theta_0 + \theta^T x_a), \tag{1.6}$$

where $\Phi(z)$ denotes the standard normal cumulative distribution function. Here, $\mu_a(\theta)$ is the conversion rate for an arm $a \in \mathcal{A}$, with x_a denote the covariate vector for the four design factors. By assuming such a model, the initial problem of learning the 625 independent arms (high-dimensional) reduces down to the more manageable problem of learning $1 + (5 - 1) * 4 = 17$ regression parameters in (1.6) (lower-dimensional). In other words, if one assumes a strictly additive structure between the covariates and conversion rates, then the size of the parameter space can be significantly reduced.

We follow the notation in [14] below. Let x_t denote the vector of indicator variables describing the characteristics of the arm played at time t. Following [14], we assume that the probability of a reward depends on x_t through a probit logistic regression model (but any other model can be used as long as we can obtain posterior sampling of parameters θ). We adopt the following two-factor interactions model to describe the interrelationship between

design factors and conversion rate response:

$$\mu_a(\theta) = \Phi(\theta^T x_a) = \Phi\left(\sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n-1} \sum_{j=1+1}^{n} \beta_{ij} x_i x_j\right). \tag{1.7}$$

While higher-order interactions can certainly be entertained, one then runs the risk of overfitting the model from the limited data available on the design space $\mathcal{A}$. Because of this, we restrict our analysis and numerical studies to the two-factor interaction model in (1.7).

The main-effect only model has the fewest model parameters for Thompson sampling to learn. Intuitively, this may produce the lowest short term cumulative regret. However, as explained in [15], [16] and [17], if the true functional relationship is rugged, then such a relationship cannot be well-captured by a main-effects model. We recommend that some two-way interactions to be included if there are enough degrees of freedom.

1.4.2 Prior specification and posterior sampling

When a large number of design factors is present, we would like to incorporate an important guiding principle, called *effect hierarchy* [18], to allow for better inference of interaction terms from limited data. The effect hierarchy principle states that lower order effects are more likely to be important than higher order effects, and that the effects of the same order likely to be equally important. One way to incorporate this within a Bayesian framework is by assigning the so-called *functional prior* [19] on effect parameters θ. Such a prior imposes effect hierarchy on θ as a prior assumption, and requires only the specification of a few hyper-parameters, which allows for easy implementation in practice. For a detailed description of this functional prior, please refer to [19]. For two-level experiments, the

functional prior has the simple form: $\beta_0 \sim \mathcal{N}(0, \tau^2)$, $\beta_i \sim \mathcal{N}(0, \tau^2 r)$ for main effects and $\beta_{ij} \sim \mathcal{N}(0, \tau^2 r^2)$ for two-way interactions. To maximize information from limited data, we adopt a fully-Bayesian perspective, and sample the hyper-parameters τ and r using Markov Chain Monte Carlo as well. While the probit logistic regression model has no conjugate priors, we can use a well-known data augmentation algorithm [20] to generate Markov Chain Monte Carlo samples from the posterior distribution $p(\theta|y)$.

1.4.3 Full algorithm

The experiment starts with a fractional factorial design with the number of configurations smaller than or equal to K. Let D_k denote the selected arms for the experiment and F_t denote the full factorial design. The generalized Thompson sampling is performed on the selected arms x_t, where $x_t \in D_k$ in each round. During each switch, we rank the arms based on $\Phi(\theta x_t)$ with $x_t \in F_t$, and the top K arms are then selected for the next round of experiments. Another advantage of assuming a model structure between arms is that, even without playing a majority of arms even once, we can still pick the best arm with some level of confidence. This can be a great advantage for our method, as we show later in numerical examples. We call our algorithm AASETS, which is an acronym for **A**ctive **A**rm **SE**lection using **T**hompson **s**ampling. The detailed algorithm is described in Algorithm 2.

Algorithm 2: AASETS

Select K arms out of 2^N for the initial design;
$Pr(Y_t = 1) = \Phi(\theta^T x_t)$ where the model matrix includes main effects and two-way interaction
The prior distribution are $\beta_0 \sim \mathcal{N}(0, \tau^2)$,$\beta_i \sim \mathcal{N}(0, \tau^2 r), i = 1, ..., p$ and $\beta_{ij} \sim \mathcal{N}(0, \tau^2 r^2), i = p+1, ..., p+\binom{p}{2}$; $\tau \sim \mathcal{N}(0, 10^2)$ and $r \sim \mathcal{U}(0, 1)$
foreach $s = 1, 2, ...S$ **do**
  foreach $t = 1, 2, ..., t$ **do**
    For each arm $i = 1, ..., K$, generate draws of $\theta_i(t)$ using data augmentation algorithm and functional prior;
    Play arm $i(t) := argmax_i \left(\theta_i(t)^T x_t\right)$ where $x_t \in D_t$ and observe reward γ_t;
  Rank the arms based on $\theta^T x_t$ where $x_t \in F_t$, select the top K arms for the next iteration

1.5 Numerical study

1.5.1 Simulation set-up

Next, we conduct some simulation studies to explore the performance of the proposed AASETS algorithm. We will use the simulation framework in [21], where effect parameters are populated from 113 published experiments across different engineering domains. The probit logistic regression model in (1.7) is used for generating the true conversion rates. First, main effects are simulated from the following distribution:

$$\beta_i|\gamma_i \sim (1-\gamma_i)\mathcal{N}(0, 1^2) + \gamma_i \mathcal{N}(0, 10^2).$$

Here, $p(\gamma_i = 1) = 1 - p(\gamma_i = 0) = P_1 = 0.41$ is the probability that the main effect β_i is significant. Next, two-way interaction effects are simulated from the following distribution:

$$\beta_{ij}|\gamma_{ij} \sim (1-\gamma_{ij})\mathcal{N}(0, 0.278^2) + \gamma_{ij} \mathcal{N}(0, 2.78^2),$$

where $p(\gamma_{ij} = 1) = 1 - p(\gamma_{ij} = 0) = P_2$ is the probability of an interaction effect being significant. Here, we want to incorporate the principle of *effect heredity* [18], which states that an interaction effect is active only when one or both of its parent effects are active. (For an interaction effect AB, we call the two main effects A and B as its parents). This effect heredity can be incorporated by sampling probability P_2 as:

$$f(P_2) = \begin{cases} 0.33, & \text{if both parent effects are significant.} \\ 0.045, & \text{if one of the parent effects are significant.} \\ 0.0048, & \text{If none of parent effects are significant.} \end{cases}$$

Finally, we simulate three-way interactions from the following distribution:

$$\beta_{ijk}|\gamma_{ijk} \sim (1 - \gamma_{ijk})\mathcal{N}(0, 0.137^2) + \gamma_{ij}\mathcal{N}(0, 1.37^2),$$

where $p(\gamma_{ijk} = 1) = 1 - p(\gamma_{ijk} = 0) = P_3$. To incorporate heredity, we sample probability P_3 from the distribution

$$f(P_3) = \begin{cases} 0.15, & \text{if all three parent effects are significant.} \\ 0.067, & \text{if two of the parent effects are significant.} \\ 0.035, & \text{if one of the parent effects is significant.} \\ 0.012, & \text{if no parent effects are significant.} \end{cases}$$

The data are then transformed to probabilities using the link function $\Phi(X^T\beta/k)$, where k is a tuning constant which ensures the values of Φ values between 0.2 and 0.8.

1.5.2 Experiments without budget constraint

We first study the scenario where all the configurations can be observed and chosen. Here, we are mainly interested in studying how the magnitude of interactions affects the outcomes. We compared the following three schemes:

1. The first set-up is the binomial bandit Thompson sampling which assumes no relationship between the arms.

2. The second set-up is the fractional factorial bandit with main effect only and non-informative prior.

3. The third set-up is the AASETS algorithm with no switches, which includes all the main effects and two-way interactions with the functional prior [19].

The algorithm is updated in batches, each with 100 samples. For each simulation, 100 updates are made, therefore the total sample size is 10,000. We then record the corresponding cumulative regret. We start the experiment with $64(=2^6)$ arms. The results are based on the average of 10 simulations, and is summarized in Figure 1.1. Here, the x axis corresponds to the number of batch updates, and the y axis corresponds to the cumulative regret defined in (1.3).

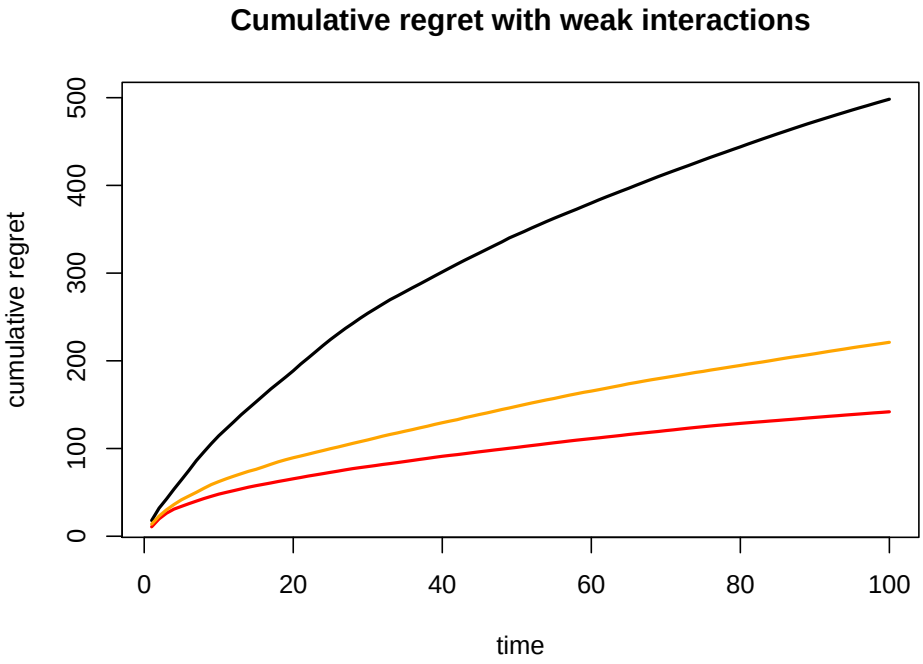

Figure 1.1: *Weak interactions*
vanilla TP, main-effect-only TS, AASETS

From the above simulation, we can see that the main-effect model actually performs better than models which incorporate interactions. This can be explained by the simulation set-up in section 5.1, which assumes that the magnitude of significant main effects are larger than the magnitude of significant two-way interactions. Indeed, such a result is not surprising, since a main-effect model has been shown to enjoy excellent optimization performance when the underlying surface is nearly additive (this is the so-called *marginal-conditional requirement (MCR)* in [15]).

When we change the magnitude of two-way interactions to be the same as the main effects, i.e. $\beta_{ij}|\gamma_{ij} \sim (1-\gamma_{ij})\mathcal{N}(0,1^2)+\gamma_{ij}\mathcal{N}(0,10^2)$, the models which include interactions now show superior performance over the main-effect only model. This is shown in Figure 1.2.

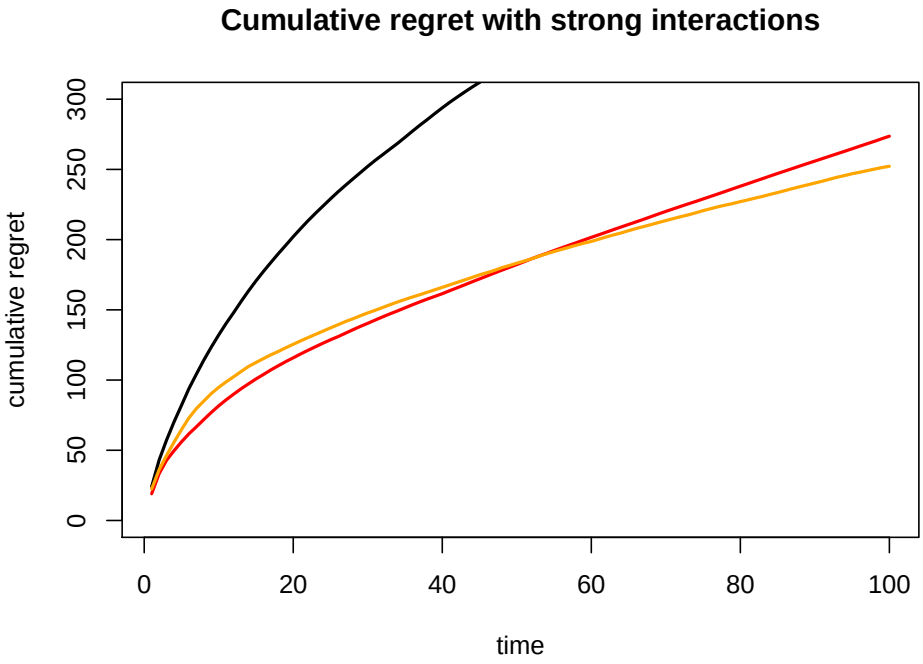

Figure 1.2: *Strong interactions*
vanilla TP, main-effect-only TS, AASETS

From the above comparison, it is not hard to see that a low-order model assumption makes the algorithm converges much faster compared to binomial bandit Thompson sampling. Moreover, if the *MCR* condition holds [15] (model includes only main effects and some minor interactions), the main-effect only model performs on average slightly better than models which include interactions. Intuitively, the main-effect only model has the smallest number of parameters to learn, which makes learning such parameters easier with limited data. However, when large interactions are present, the main-effect only model can be inferior.

1.5.3 Experiments with budget constraint

In real world online testing problems, the company cannot afford to test *all* the combination settings at the same time. Suppose there are 10 factors in total and each factor has two levels, but in each round only $16(=2^4)$ arms can be selected and compared. We assume the experimenter can switch from old arms to new arms 5 times based on the criterion list in benchmarks. As mentioned in an earlier illustrating example, the switching of arms typically happens at the end of the day, when traffic is at its lowest. Within each experiment, 50 updates are performed and each update is in a batch of 100. For all the figures below, the *cumulative regret* is defined as (1.5).

We compared our proposed method with the following three benchmarks, which are commonly used in the industry.

1. Benchmark 1: The first baseline uses "vanilla" Thompson sampling in each phase, and the best arm returned by Thompson sampling is deployed for the rest of the session. Here, "best arm" refers to the arm which has the largest posterior probability of being the optimal arm from Thompson sampling.

2. Benchmark 2: The second baseline is another popular industry method. In each round of the experiment, K arms are randomly selected from the arm pool and Thompson sampling is performed within the selected arms. Then, at the end of each round, the sub-optimal arms (arms with less than 5% posterior probability of being the best arm) will be discarded, and new arms will be randomly added to the candidate arm set.

3. Benchmark 3: The third baseline performs Thompson sampling within each round

of the experiment. At the end of the round, the top K arms are selected based on their posterior probabilities based on Thompson sampling. Note that this method will always select the top arms, and it may result in discarding all of the previous selected arms.

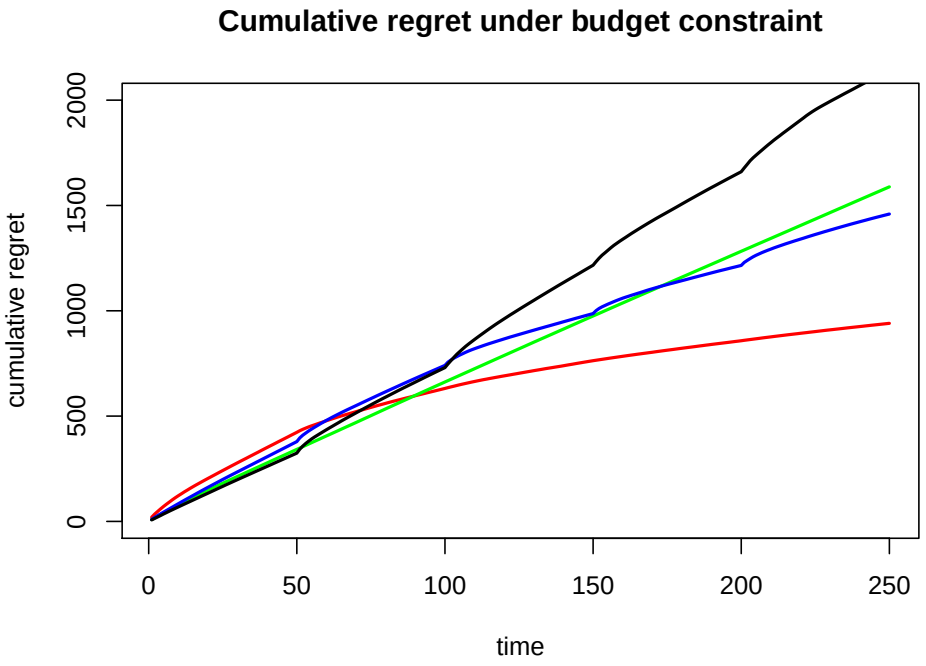

Figure 1.3: *Cumulative regret*
AASETS, benchmark 1, benchmark 2, benchmark 3

Figure 1.3 shows the cumulative regret of the four methods over the experimental period. At the end of the experiment, AASETS has much lower cumulative regret over the other three methods, which suggests our proposed algorithm outperforms all three industry benchmarks by a large margin. At round 1 (time 0 - 50), AASETS is not significantly better than the other three benchmarks. This is expected since all three benchmarks use "vanilla" Thompson sampling at round 1, which assumes no dependency between arms. The total number of arms is 16, which is the same as the number of unknown parameters in AASETS. Because of this,

the faster convergence of AASETS (from a reduced number of model parameters) cannot be exploited at the start of the experiment. As the experiment progresses, however, our method shows significantly lower cumulative regret. This can be attributed to the "switching" procedure of our algorithm, which uses the fitted low-order interactions model to actively add promising new arms and remove poor performing arms.

1.5.4 More simulation settings under budget constraint

Next, to study our algorithm under more general settings, we compare the proposed AASETS method with the earlier three industry benchmarks under the following modifications:

1. Increasing the total number of arms from $N = 2^8$ to $N = 2^{15}$.

2. Incorporating "strong" and "weak" two-way and three-way interactions.

3. Changing the arm budget constraint (i.e., the number of arms an experimenter can select in each round) to 16 arms, 32 arms and 64 arms.

The following subsections discuss our results under the above simulation modifications.

Increasing the total number of arms

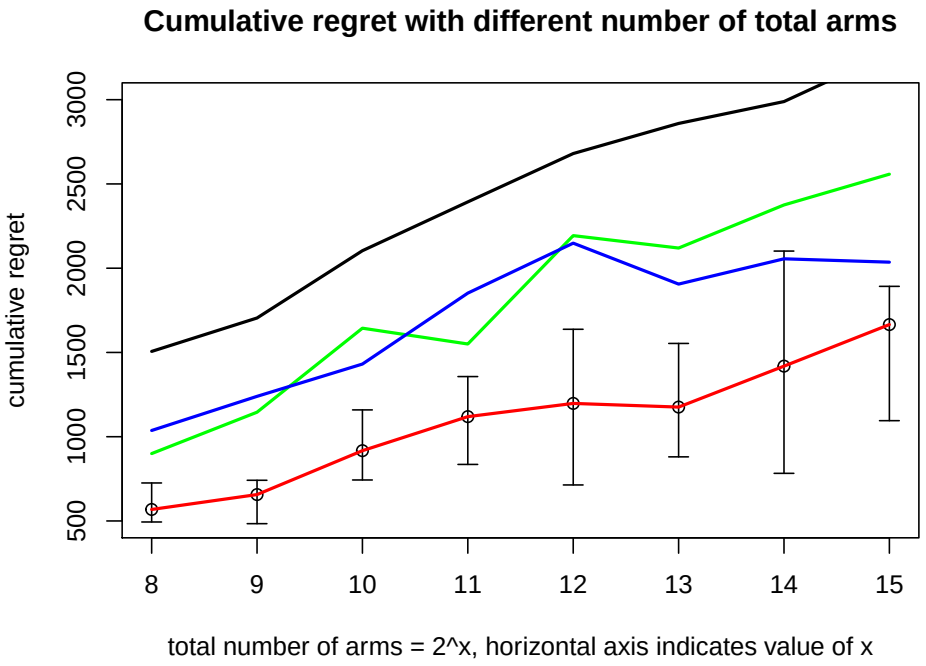

Figure 1.4: *Total regrets*
AASETS, benchmark 1, benchmark 2, benchmark 3

Figure 1.4 shows the cumulative regret for the four methods at the end of the experimental period, with total arms N ranging from 2^8 to 2^{15}. Here, AASETS gives noticeably lower regret compared to the other three benchmarks. Moreover, from the slopes of the lines in Figure 1.4, we see that AASETS has a smaller slope compared to Benchmarks 1 and 2, and a comparable slope to Benchmark 3. This suggests that the improvement of our method over industry benchmarks grows larger as the total number of arms increases (i.e., as the design space of arms grows larger). This is again expected, since our method allows us to infer information on unobserved arms using the experimental data collected from a small number of observed arms.

"Strong" and "Weak" interactions

Next, we study the performance of AASETS under weak and strong two-way and three-way interactions. Following Section 4.1, weak interactions are simulated from the distribution:

$$\beta_{ij}|\gamma_{ij} \sim (1-\gamma_{ij})\mathcal{N}(0, 0.278^2) + \gamma_{ij}\mathcal{N}(0, 2.78^2)$$

and

$$\beta_{ijk}|\gamma_{ijk} \sim (1-\gamma_{ijk})\mathcal{N}(0, 0.137^2) + \gamma_{ij}\mathcal{N}(0, 1.37^2).$$

Strong interactions are simulated from the distribution:

$$\beta_{ij}|\gamma_{ij} \sim (1-\gamma_{ij})\mathcal{N}(0, 1^2) + \gamma_{ij}\mathcal{N}(0, 10^2)$$

and

$$\beta_{ijk}|\gamma_{ijk} \sim (1-\gamma_{ijk})\mathcal{N}(0, 1^2) + \gamma_{ij}\mathcal{N}(0, 10^2),$$

i.e., two-factor and three-factor interaction effects are assumed to have the same magnitude as main effects. Note that for the weak interactions model, three-factor interactions β_{ijk} have a smaller magnitude than two-factor interactions β_{ij}, whereas for the strong interactions model, they have the same magnitudes.

Figure 1.5 shows the cumulative regret as a function of experimental time, for each of the

Comparison under "weak" (top) and "strong" (bottom) interactions

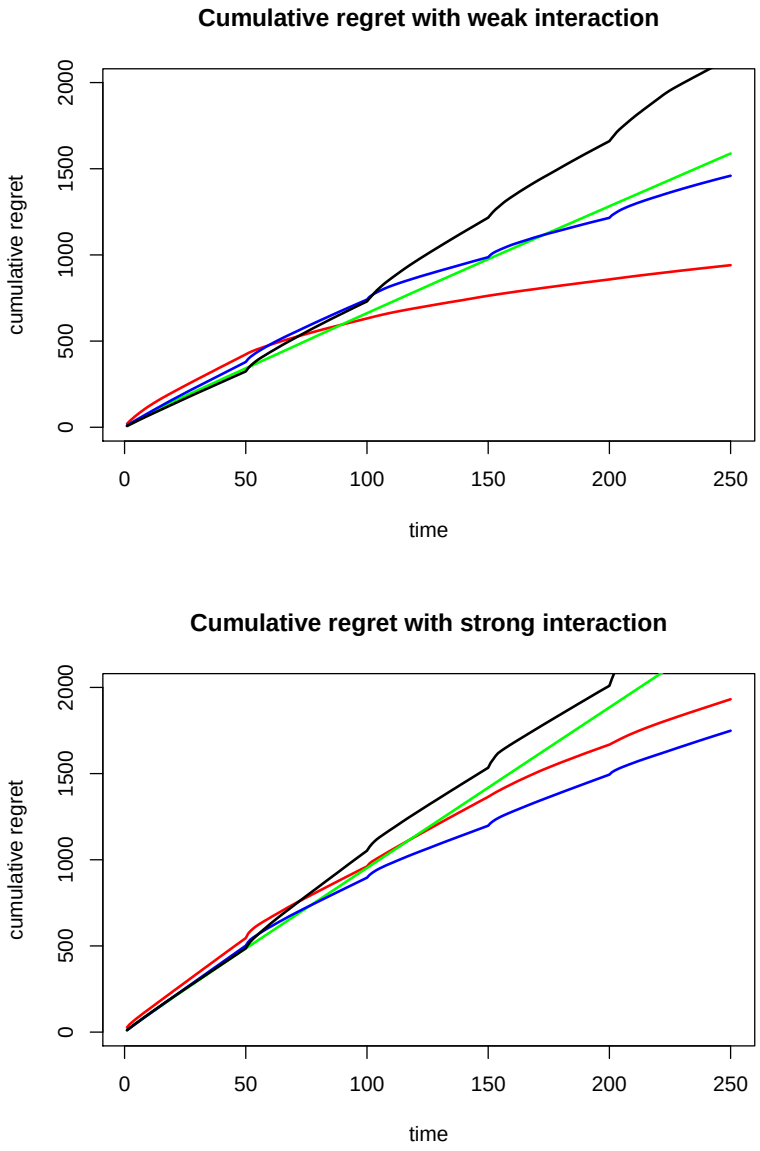

Figure 1.5: *Weak interactions vs strong interactions*
AASETS, benchmark 1, benchmark 2, benchmark 3

four methods. Under weak interactions, the AASETS method again outperforms the existing three benchmarks by a large margin. Under strong interactions, however, AASETS has a slight disadvantage compared to Benchmark 2. One explanation for this is that, when the true underlying surface is highly non-additive with strong interactions, a model-based approach

for optimization can perform poorly, since it is difficult to learn these strong interactions well with limited data. One approach is to employ a hybrid rank- and model-based approach to optimization (see, e.g., [15]), but we defer this to future work.

Changing the arm budget constraint

Lastly, we study the performance of our method under different budget constraints (other parameter settings are the same in section 4.3). Recall that a budget of $K = 16$ arms is used as the baseline setting in the earlier Figure 1.3. We now test a larger budget size of $K = 32$ and $K = 64$.

Figure 1.6 shows the cumulative regret of the four methods under this modification. We see that, while AASETS still maintains an improvement over the other three methods, the improvement gap over Benchmark 1 grows smaller as the arm budget size grows larger. This is again intuitive, since the key advantage of AASETS is that it can infer information on unobserved arms from experimental data observed on a small budget of arms. When this budget grows large (relative to the total number of arms), the value gained from this advantage decreases. In this sense, AASETS is expected to be most effective when there is a tight arm budget constraint.

1.6 Real world experiment

Finally, we apply the proposed AASETS method on a real-world bandit dataset. While it would be nice to test the method on a practical e-commerce application, such data is typically not made available as publicly-available datasets. We will instead use a dataset from the Kaggle competition, called the **Forest Cover Type Dataset**. This dataset can be

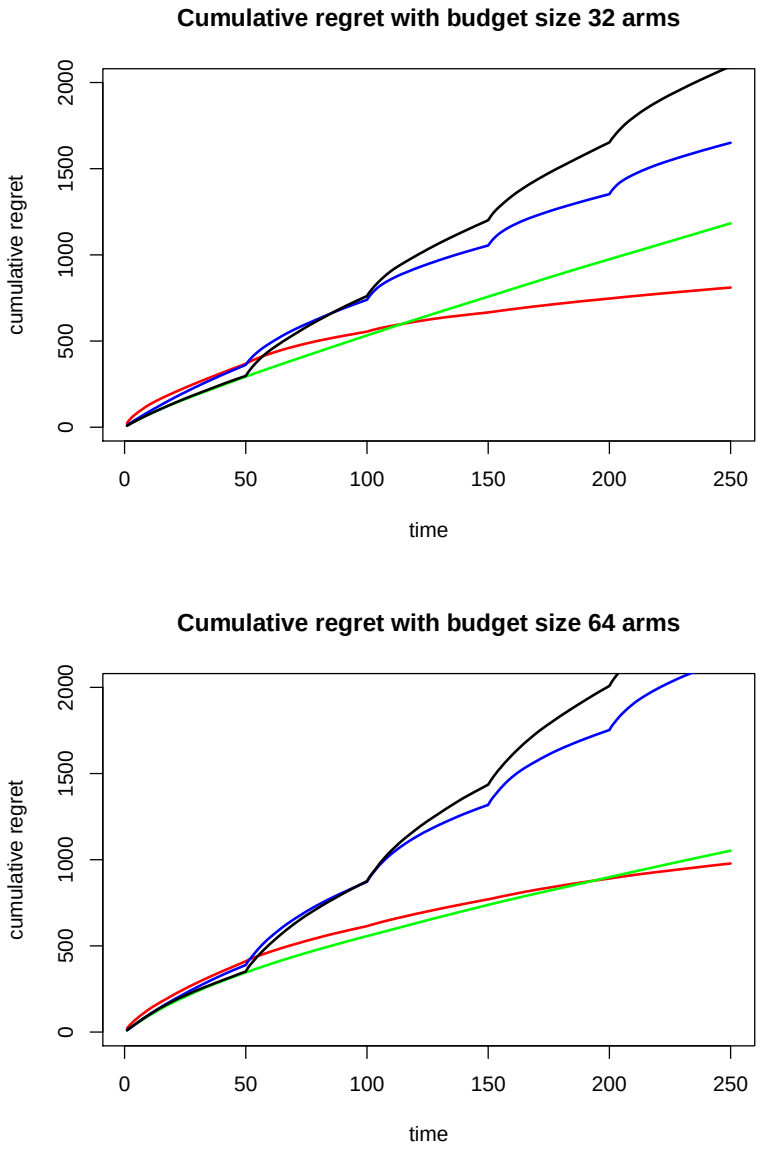

Figure 1.6: *32 arms vs 64 arms*
AASETS, benchmark 1, benchmark 2, benchmark 3

obtained from *https://www.kaggle.com/uciml/forest-cover-type-dataset*.

To fit our problem set-up, we partition the initial dataset into 512 clusters using k-means clustering algorithm [22] on its covariates. Each cluster i has been treated as an arm, and its response is calculated as the average of all responses from cluster i. The 9 covariates from the dataset are then treated as 9 factors in our experiment. We still assume the budget size is

$K = 16$, with 50 batch updates in each round.

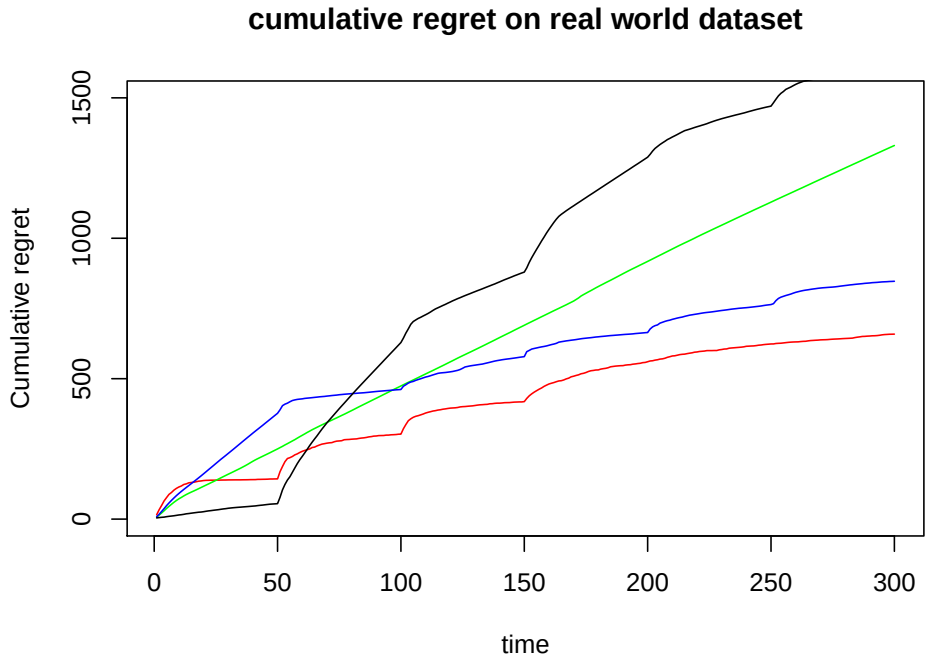

Figure 1.7: *Real-world example cumulative regret*
AASETS, benchmark 1, benchmark 2, benchmark 3

Figure 1.7 shows the cumulative regret for the four methods, as a function of experimental time. We can see that AASETS again outperforms the other three industry benchmarks, by achieving much lower cumulative regret at the end of the experimental period. Indeed, AASETS is the only method (of the four tested) which achieves a "plateauing" behavior for cumulative regret L over time. Given that this cumulative regret is the sum of regret L_t over all previous time periods t (see equation (1.3)), this implies that regret L_t is consistently decreasing and converging to zero at the end of the experimental period. This desired behavior is made possible through the fitted low-order interactions model, which allows an experimenter to gain insight on reward probabilities for a large proportion of unplayed arms, using the experimental data from a limited number of observed arms.

1.7 Summary and future work

The problem of arm budget constraint (where an experimenter has a limited budget of arms which can be tested in a given time period) is a widely-encountered problem in real-world website design. But to our knowledge, this problem has not been directly addressed in the multi-armed bandit literature. We propose in this paper a new method, called Active Arm Selection using Thompson sampling (AASETS), which performs active arm selection and traffic allocation in an online setting, under such an arm budget constraint. The key novelty of AASETS is the use of a low-order interactions model to learn dependencies between arms on a large, factorial design space. Using this fitted model, AASETS allows an experimenter to (i) adaptively add good arms and remove bad arms from experimentation – a procedure we call "switching", and (ii) leverage conversion results from all arms for effective traffic allocation. We then show that AASETS outperforms several industry benchmark methods under the assumed arm budget constraint, both in simulations and a real-world example.

Motivated by the encouraging results from this work, there are several interesting avenues for future work. One such direction is the incorporation of a hybrid scheme with AASETS which adaptively integrates rank- and model-based optimization. As explored in [16], [17], and more recently in [15], a rank-based optimization method may outperform a model-based method when (i) the objective function f is rugged, and (ii) one has limited data on f. This development will allow for a more robust MAB method which can adaptively exploit within-arm dependencies from data. Another interesting direction is the extension of AASETS for more general class of reward distributions (beyond binominal bandits), which will allow for greater applicability of our method

CHAPTER 2

CFO: CONDITIONAL EFFECT BASED FUNNEL TESTING FOR CONVERSION RATE OPTIMIZATION

2.1 Abstract

Conversion rate optimization has become more important because of the rapid growth of e-commerce revenue. Traditional conversion rate optimization, including AB testing and multivariate testing, tends to isolate factors and treat them the same regardless of their positions in the web system. [23] proposed a new method called funnel testing, which can study factors main effect and interaction based on the conversion funnel they belong to using directed graph. In this paper, we will discuss a new framework, called CME based funnel testing, where factors effects and level settings are analyzed and optimized based on their position on the webpage. We called the new approach **CFO**. The new approach has better interpretability of the factorial effect and achieves better result in conversion rate optimization.

2.2 Introduction

The transaction of buying and selling things on-line is called **E-commerce**. In the information technology era, e-commerce is claiming a bigger share of commerce and has become an important source of revenue for many companies. In Internet marketing, conversion optimization, or conversion rate optimization (CRO) is a system for increasing the percentage of

visitors to a website that convert into customers[1], or more generally, take some desired actions on a web page. It is commonly referred to as **conversion rate optimization**, or **CRO**. CRO has become a hot research topic in E-commerce and it has been extremely important in large IT and retail companies such as Microsoft, Amazon, and Walmart. Conversion rate optimization sometimes is even more important for small online business and start-ups since the entire business may depend on the conversion rate.

Two approaches are currently popular in CRO. The first one, A/B testing, which compares two or more versions of the same factor: the original version and the proposed new versions. Classical Hypothesis testing such as student's T-test is used to assess the difference and the best version is chosen as the design of the web page in the future. A/B testing is the most commonly used CRO technique because of it's simplicity. However, customer's conversion rate rarely depends only on one factor. For example, suppose the customer wants to buy products on Amazon, the customer usually start with the Amazon's homepage, then they may go through Product Listing Page, Product description Page, Checkout page etc... With so many other pages involved, studying just one factor in order to maximize the conversion can be an oversimplification, and sometimes misleading. [18] has also shown that this so called "one-factor-at-a-time" approach is sub-optimal compared to fractional factorial design. The second approach, multivariate testing (MVT), where multiple factors are studied in one experiment. MVT is implemented with fractional or full factorial designs, and models may or may not fitted to the conversion rates with respect to the factors. Its optimization is done either by choosing the optimal level settings that achieve the best conversion rate without fitting a model, which is called model-free MVT, or based on a regression model fitted to the conversion rate data and the factors[18], which called model-

based MVT. Model-free MVT achieves great optimization accuracy if all the versions of the website are tested and each version receives enough traffic, but it offers no information to the experimenter about the relationship between the conversion rate and input factors, which may not give the experimenter confidence in the optimization result. Model-based MVT can deal with multiple factors efficiently using fractional factorial design and variable selection techniques, but it does not differentiate the factors based on where they are on the pages of the website. This can lead to misleading results. Consider, for example, two conversion funnels A and B, where A consists of landing page, product description page, and checkout page; B consists of landing page and checkout page only. Suppose one of the factors of interest is the color of the checkout button. Suppose, in the absence of the product description page, red color on the checkout button gives the highest conversion rate. However, on the product description page, there is a red advertisement which has a similar size of the checkout button. Then the customer comes from conversion funnel A may be unwilling to click on the red checkout button due to the advertisement. Also, interaction effect may not be meaningful for the practitioner to optimize their website. We would like the method to be able to distinguish the color effect based on the conversion funnel user went through and optimize it's setting accordingly and offers interpretability for the practitioner so they can trust their result.

Define the series of pages the visitors go through until a possible conversion as a *conversion funnel*. [23] proposed a new framework called "funnel testing" that differentiates the factors based on the conversion funnel they belong to. However, the proposal has some deficiencies that make it less useful in practice. First, it distinguishes the same factor on different conversion funnel by making extra copies of the factor. As the structure of the

website become complex, the number of factors can easily become untraceable. Moreover, The proposed funnel testing framework fail to take into account the funnel location each factor is on. For example, when customer going through the funnel A $\rightarrow$ B, the factorial effect of page B on customer is based on the condition that customer already sees the factor on page A, we would like the framework to be able to distinguish between A $\rightarrow$ B and B $\rightarrow$ A, we therefore propose the new framework for conversion rate optimization, called CME-based funnel testing, that uses conditional main effect to take into account the position information each factor possess.

The idea of CMEs was first introduced in [24] as a way to disentangle effects which are aliased in a designed experiment. Ever since the founding work by [25], it has been widely known and accepted that aliased effects in two-level regular designs cannot be de-aliased without adding more runs. A result by Wu in his 2011 Fisher Lecture showed that aliased effects can sometimes be de-aliased using a new framework based on the reparametrization of the aliased effects into CMEs to allow for selection of the correlated conditional effects. The analysis methodology for designed experiments is further developed in [26]. In this paper, we take advantage of the conditional structure of the conditional main effect (CME) to model the sequential structure of the website system. This new framework can lead to models that are more interpretable and more accurate in conversion rate optimization.

An ideal conversion rate optimization framework must have three properties: Interpretability, optimization accuracy and simplicity for users to use. We demonstrate that CME-based funnel testing is the only framework that has all three properties. The organization of the paper is as follows. In Section 2, we review the funnel testing presented by [23] and introduce the notation of CME-based funnel testing, in Section 3, we discuss

the new proposed CME-based funnel testing, in Section 4, three examples are given to demonstrate the effectiveness and serve as a guideline for practitioners to better understand the framework, In Section 5, numerical studies are conducted to show that CME-based funnel testing performs well in terms of finding the optimal factorial settings to maximize the conversion rate. Concluding remarks are given in the last section.

2.3 Background on funnel testing

Funnel testing is first proposed by [23] as a new framework for conversion rate optimization that treats factors differently based on the conversion funnels they belong. [23] used directed graph to study the Internet where all the web pages in a conversion system are viewed as vertices of the graph. If there is a hyper-link on page X referring to page Y, draw a directed edge from X to Y. We will follow this convention in this paper to represent the Website.

After representing the conversion system as a directed graph, the next step is to identify all the conversion funnels. Conversion funnel is defined as a series of pages that a visitor has gone through before making a conversion. The method is then build a statistical design based on the number of factors in the conversion system. The testing should be carried out online with equal traffic distribution to each of the factorial combination settings. For each conversion funnel CF_i, the significant factors are first identified and a linear model is build based on the significant factors.

$$CR_i = \beta_{i0} + \beta_{i1}p_1 + \beta_{i2}p_2 + ... + \beta_{ij}k_1 + \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

where page $p, k \in CF_i$.

When analyzing the experiment, the author claims if the significant factors, for example p_1 happens belongs to multiple conversion funnels i, j, the method needs to check the path previous of page p between funnel i and funnel j, if the path is different, an additional copy of p_1 is made in j as j' and use j' as factor in the linear model of CR_j. We can see this approach is not feasible when the website system grows large. In today's website system each web page is shared by at least tens or even hundreds of conversion funnels and we cannot afford to allow the number of factors to grow as the complexity of the system grows.

After the model for each conversion funnel CF_i is defined, overall conversion rate is defined as

$$CR_t = \sum_{i=1}^{n} w_i * CR_i = \sum_{i=1}^{n} \beta_{i0} + \sum_{i=1}^{n} \beta_{i1} p_1 + \sum_{i=1}^{n} \beta_{i2} p_2 + ... + \sum_{i=1}^{n} \beta_{ij} k_1$$

The last step is to perform optimization to find the optimal level settings of each factor in the system. Again we can see the optimization is done on all the variables and all the variables are treated identically. Intuitively we can see this method is not efficient and fail to take advantage of the conditional dependent structure the Internet possesses.

The proposed new framework uses the concept of conditional main effect(CME) to describe the conditional dependence structure of the website. CME was first introduced by [24] and the analysis strategies for designed experiments is further developed in [26]. Suppose we have two factors A and B and each has two levels, denoted as "-" and "+", the conditional main effect of A given B at level $+$ is defined as

Definition 2.3.1.

$$CME(A|B+) = \bar{y}(A+|B+) - \bar{y}(A-|B+)$$

A	B	$A \mid B+$	$A \mid B-$	$B \mid A+$	$B \mid A-$
+1	+1	+1	0	+1	0
+1	-1	0	+1	-1	0
-1	+1	-1	0	0	+1
-1	-1	0	-1	0	-1

Table 2.1: *Model matrix for the MEs A and B and their corresponding CMEs*

where $\bar{y}(A+|B+)$ denotes the average of y_i values with both A and B at the $+$ level and $\bar{y}(A-|B+)$ is similarly defined.

Its contrast vector is defined by

Definition 2.3.2. The conditional main effect contrast of A given B at level $B+$, denoted as $A|B+$, quantifies the covariate vector $A|B+ = (A_1, A_2, ...A_n)|B+$ where

$$A_i|B+ = \begin{cases} A_i & \text{if } B = +1 \\ 0 & \text{if } B = -1 \end{cases}$$

for $i = 1, ..., n$

2.4 Analysis of CME funnel testing

2.4.1 System with one conversion funnel and 2-level factors only

Consider a simple conversion funnel A $\rightarrow$ B $\rightarrow$ C (figure 1). We can assume A is the homepage, B is the product description page and C is the checkout page. Assume there is one factor on each page denoted as **a, b, c** and each one has two levels only. For example, factor **a** is the background color for homepage, in which **a-** stands for orange color and **a+** stands for red color. Factor **b** represents the font on product description page in which **b-**

means "small" font and **b+** means "large" font; factor c is the shape of the checkout button on checkout page in which **c-** represents rectangular and **c+** represents oval shape of the button. Then we define the CME-based rules as follows:

- Rule 1: Keep the main effects of the landing page.
- Rule 2: If there is an arrow from page i to page j, exchange the main effect j and interaction of ij with conditional main effect $j|i+$ and $j|i-$. Then perform variable selection to select the important main effects and conditional main effects.
- Rule 3: assume there is no interaction between factor i, j if there is no edge between i, j.

Here, "-" and "+" stand for the two level settings each factor has. Consider the simple conversion funnel defined in figure 1,

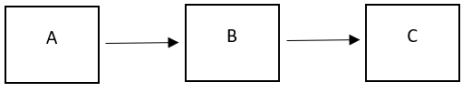

Figure 2.1: *Conversion system with one conversion funnel*

Where the effects of interest are main effect A,B,C, interaction effects AB,BC,AC and ABC. Using the rule we defined, we keep the main effect of A since A is the landing page; since there is an link from page A to page B and from page B to page C, we exchange B, AB with $B \mid A+$ and $B \mid A-$ and exchange C,BC with $C \mid B+$ and $C \mid B-$. Since there is no directed edge between A and C, we assume AC has no interaction or conditional effect. So the variable left in the model is A, $B \mid A+$, $B \mid A-$, $C \mid B+$, $C \mid B-$;

Rule 2 is based on the most important selection rule proposed by [26]. It is based on two simple mathematical identities

$$(A|B+) = \frac{1}{2}(A + AB)$$

$$(A|B-) = \frac{1}{2}(A - AB)$$

From the above equation, the CME $(A \mid B+)$ can be viewed as an average of the main effect for A and the interaction effect for AB; a similar interpretation holds for the CME $(A \mid B-)$. Rule 1 of [26] replaces a selected main effect A and 2-factor interaction AB with either (a) The CME $A \mid B+$ if the effect magnitudes are similar and A and AB have the *same* sign. Notice if they meet the above criteria, $A \mid B-$ will be 0 and have no usefulness to be included in the model.Or (b) CME $A \mid B-$ if the effect magnitudes are similar and A and AB have the *opposite* sign. It is clearly seen that the conditional effect also takes into account the *sequential* nature of the web page system.

Rule 1 and Rule 3 are based on the Markovian property of a directed acyclic graph, which is assumed for two reasons: First, to reduce the number of variables that potentially affect the conversion rate. Second, customer's behavior on a specific website is usually affected by only the current web page or the previous couple of pages.

2.4.2 System with multiple conversion funnels and 2-level factors only

Consider graph in figure 2, where there is a conversion funnels from page A to page C directly.

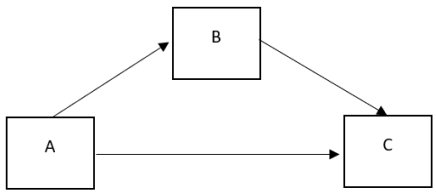

Figure 2.2: *Conversion system with two conversion funnels*

If there is multiple conversion funnels in the system, we use part 1 to analyze each conversion funnel. For this example, funnel one from A $\rightarrow$ B $\rightarrow$ C will have variable A, $B|A+$, $B|A-$, $C|B+$, $C|B-$ and funnel two from A $\rightarrow$ C will have variable A, $C|A+$ and $C|A-$. We combine them together in the end to form our objective function to optimize.

We can see how the CME-based funnel differentiate same factor from different channels by conditional on the previous factor in the funnel.

2.4.3 System with multiple conversion funnels which contains factors with 3 or more levels

The factor of interest may have multiple levels that need to be compared. For example, the background color of a homepage can be orange, red or brown. When factors have three levels, the conditional main effect between factors has not been studied before. We propose to use the following framework to model the conversion system. For a quantitative factor, say A, [18] suggest use A_l, A_q (linear and quadratic effect) for A's main effect. For a qualitative factor D, which is more common in online testing, we can use D_l, D_q if D_q is interpretable; otherwise, select two contrast from D_{01}, D_{02}, D_{12} as D's main effect, where the contrast is defined as follow where each of the three contrasts compares only two out of

the three levels.

$$D_{01} = \begin{cases} -1 & \text{for level 0} \\ 1 & \text{for level 1} \\ 0 & \text{for level 2} \end{cases}$$

$$D_{02} = \begin{cases} -1 & \text{for level 0} \\ 0 & \text{for level 1} \\ 1 & \text{for level 2} \end{cases}$$

$$D_{12} = \begin{cases} 0 & \text{for level 0} \\ -1 & \text{for level 1} \\ 1 & \text{for level 2} \end{cases}$$

In terms of conversion rate optimization, the conditional main effect should be studied separately at each level of its parents factor. For factor B's level depend on A, we propose analyzing the effect of the factor B at the i_{th} level of A, denoted by $B|A_i$. If B is quantitative with more than two levels, then the linear, quadratic, etc. effects of B at the i_{th} level of A, denoted by $B_l|A_i$,$B_q|A_i$ are analyzed. If B is qualitative, we can pick the two contrast B_{01} and B_{02}, where B_{ij} denotes the contrast between level i and j level j of factor A.

The analysis framework proposed by [26] restrict the search to orthogonal models, we need to relax this assumption for the funnel testing because of the number of factors that will be in the model. Instead we using multiple regression techniques, such as stepwise

regression or subset selection procedure to identify a suitable model. The strategy can be summarized as follow:

- For a factor on the landing page, say A, if A is quantitative, use A_l and A_q for A's main effect; If A is qualitative, use A_l and A_q if A_q is interpretable; otherwise, select two contrasts from A_{01}, A_{02}, A_{12} for A's main effect;

- If factor B follows factor A on a web page system, We propose analyzing the effect of the factor B at the ith level of A, denoted by $B|A_i$ for each i. If B is quantitative with more than two levels, then the linear, quadratic, etc effects of B at the ith level of A, denoted by $B_l|A_i$,$B_q|A_i$ are included in the model. If B is qualitative, we can pick the two contrasts such as B_{01} and B_{02}, where B_{ij} denotes the contrast between level i and level j of factor B.

- Using the contrasts define in 1-2 for all the factors as candidate variables, perform a stepwise regression or lasso-like model selection procedure to identify a suitable model.

2.5 Examples

In this section, we use three simulated examples to demonstrate the analysis strategy and to show the advantage of the CME-based funnel testing not only on better modeling result, but also on the interpretability of the result .

A	B	C
-1	-1	-1
-1	-1	+1
-1	+1	-1
-1	+1	+1
+1	-1	-1
+1	-1	+1
+1	+1	-1
+1	+1	+1

Table 2.2: *Design matrix for toy example*

2.5.1 Example 1: Toy example

Consider a website system only contains three pages: a landing page i.e. homepage, a product description page and a converting page. The system is defined as figure 2.3

Figure 2.3: *Conversion system with one conversion funnel*

We assume one factor per page and use A,B,C to represent the factors on landing page, product description page and converting page and each has two levels. The system only contains one conversion funnel. We use 2^3 full factorial design and the design matrix is as table 2.2:

For each simulation, we first choose a version of the conversion system from the 8 candidates in the design table with equal probability, and then simulated the visitors'

Functions for decision probabilities
$t_1 = 0.5 - 0.1A$
$t_2 = 0.5 + 0.2B - 0.25AB$
$t_3 = 0.38 - 0.1C - 0.15BC$

Table 2.3: *Probability distribution*

behavior according to the probability distribution given in table 2.3. The visitor always starts with the landing page, then he/she has two choices :go to product description page with probability t_1 or leave the system with probability $1 - t_1$. If the customer decide to visit the product description page, he/she then have two choices guided by t_2: Visiting the conversion page or leave the system. Note that,although the choices are made on page 2,the probability t_2 are functions of both factor A and B, because it is believed that the previous visited page will affect the visitor's behavior thereafter. Once the customer in the conversion page he/she can choose to convert by probability t_3 or leave the system with probability $1 - t_3$.

Based on the analysis strategies we defined in section 3.1, the variable left in the model is A, $B \mid A+$, $B \mid A-$, $C \mid B+$, $C \mid B-$, We list the design matrix ,their corresponding conversion rate and the covariates for CMEs in table 2.4

To demonstrate the superiority of our model, we compare the CME-based funnel testing to funnel testing proposed by [23]. Since there is only one funnel in the entire system, funnel testing is essentially reduced to model based multivariate testing. The first step of modeling is to make a half-normal plot combine with Lenth's method to identify significant effects. In Figure 2.4

A	B	C	Conversion Rate	$B \mid A+$	$B \mid A-$	$C \mid B+$	$C \mid B-$
-1	-1	-1	0.0121	0	-1	0	-1
-1	-1	+1	0.0102	0	-1	0	+1
-1	+1	-1	0.3323	0	+1	-1	0
-1	+1	+1	0.0701	0	+1	+1	0
+1	-1	-1	0.0859	-1	0	0	-1
+1	-1	+1	0.1094	-1	0	0	+1
+1	+1	-1	0.0987	+1	0	-1	0
+1	+1	+1	0.0267	+1	0	+1	0

Table 2.4: *Design Matrix and conversion rate*

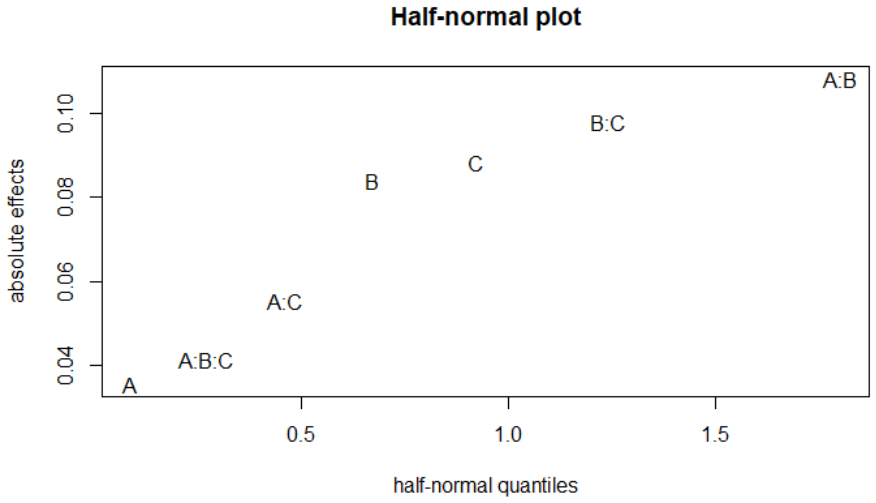

Figure 2.4: *half-normal plot*

It is not clearly which effects are the significant effects. The t-like statistics for Lenth's method are described in table 2.5.

Compare with the critical values for individual error rate(IER), only AB and BC are significant at $\alpha = 0.4$ level. Combine this result with half-normal plot and effect Hierarchy principle, the final model have four terms including main effect B,C and interaction effect between AB and BC. Therefore,the model has four terms. The R^2 value for this model is 0.861 and the adjusted R^2 value is 0.6756. The P value for the F-test is 0.1188 and the P

Effect	$\|t_{PSE}\|$
A	0.2234
B	0.6667
C	0.6717
AB	0.9675
BC	0.7644
AC	0.4635
ABC	0.3542

Table 2.5: $|t_{PSE}|$ *Values for multivariate testing*

values for the four terms are $0.07, 0.12, 0.16, 0.15$. The explicit expression of the model is

$$CR = 0.09319 - 0.05626AB - 0.04445BC + 0.03877B - 0.03906C$$

.

We now analyze the same data using CME-based funnel testing. Based on rule 1 and rule 2, variable A, $B|A+$, $B|A-$, $C|B+$ and $C|B-$ are included as our candidate variables. Stepwise regression with forward selection is then performed and $B|A-$ and $C|B+$ are added to the model. Therefore, the final model only contains two terms. The R^2 value for this model is 0.84 and the adjusted R^2 value is 0.78. The P value for the F-test is 0.01 and the P values for the two terms are $0.011, 0.018$. The explicit expression of the model is

$$CR = 0.09319 + 0.095B|A^- - 0.084C|B^+ \tag{2.1}$$

The modeling result from CME-based funnel testing is better than funnel testing on many levels. First, the R^2 value from CME-based funnel testing is higher compare to funnel testing means the selected CMEs can better explain the variance in our data compare to selected

main effects and interaction. Second, the selected CMEs are more significant compare to the main effects and interaction included in the traditional funnel testing. Moreover, the selected CMEs have good business interpretation and can help the experiment observers to better understand the interaction between different pages. For example, from model (2.1), factor B's setting should be set to "+" when A is set as "-" and C should be set as "-" when B is set as "+".

2.5.2 Example 2: System with multiple conversion funnel

Now consider a website system with two conversion funnels as defined in figure 2.5. Both conversion funnels have the same starting point(homepage) and the same conversion point(checkout page). The difference is the first funnel has one additional intermediate page,the product description page, between the homepage and conversion page.

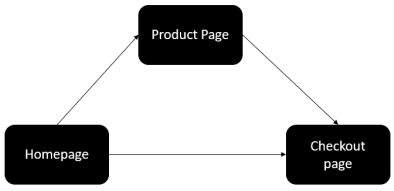

Figure 2.5: *Conversion system with more than one conversion funnel*

Like the last example, we assume one factor per page and use A,B,C to represent the factors on landing page, product description page and checkout page. The design is still a 2^3 full factorial design as in table 2.2.

The simulation is created based on the probability distribution given in table 2.6.

The visitor always starts with the landing page, then she/he can either stays in the website

Functions for decision probabilities
$t_1 = 0.5 - 0.1A$
$t_{ab} = 0.5 + 0.2B - 0.25AB$
$t_{ac} = 0.38 - 0.1C + 0.05A - 0.1AC$
$t_{bc} = 0.5 + 0.3C - 0.1BC$

Table 2.6: *Probability distribution*

A	B	C	$B \mid A+$	$B \mid A-$	$C \mid B+$	$C \mid B-$	$C \mid A+$	$C \mid A-$	Conversion Rate 1	Conversion Rate 2
-1	-1	-1	0	-1	0	-1	0	-1	0.0065	0.1109
-1	-1	+1	0	-1	0	+1	0	+1	0.0243	0.0966
-1	+1	-1	0	+1	-1	0	0	-1	0.1568	0.0864
-1	+1	+1	0	+1	+1	0	0	+1	0.4224	0.0924
+1	-1	-1	-1	0	0	-1	-1	0	0.0186	0.1120
+1	-1	+1	-1	0	0	+1	+1	0	0.1829	0.0512
+1	+1	-1	+1	0	-1	0	-1	0	0.0545	0.1287
+1	+1	+1	+1	0	+1	0	+1	0	0.1528	0.0579

Table 2.7: *Design matrix and conversion rate*

system with probability t_1 or leave the system with probability $1 - t_1$. If he/she stays in the system, we assume the customer has equal probability to visit product description page or go directly to checkout page. If the customer decide to visit product description page next, he/she has t_{ab} percent of the chance to visit checkout page and convert at checkout page with probability t_{bc}. If the customer decide to visit directly from homepage to checkout page, he/she has t_{ac} percent of the chance to convert.

Again, we compare result between CME-based funnel testing with funnel testing side-by-side. Both analysis framework analyze each conversion funnel separately. Define funnel A-B-C as funnel I and funnel A-C as funnel II. Their design matrix, conversion rate, and their CME covariates from analysis strategy 3.2 are listed in table 2.7.

We first analyze the system using funnel testing. The half-normal plot and $|t_{pse}|$ from Lenth's method for funnel 1 are given in figure 2.6 and table 2.8.

Effect	$\|t_{PSE}\|$
A	0.4225
B	1.2167
C	1.1989
AB	1.1913
BC	0.3803
AC	0.0593
ABC	0.6667

Table 2.8: $|t_{PSE}|$ *Values for multivariate testing*

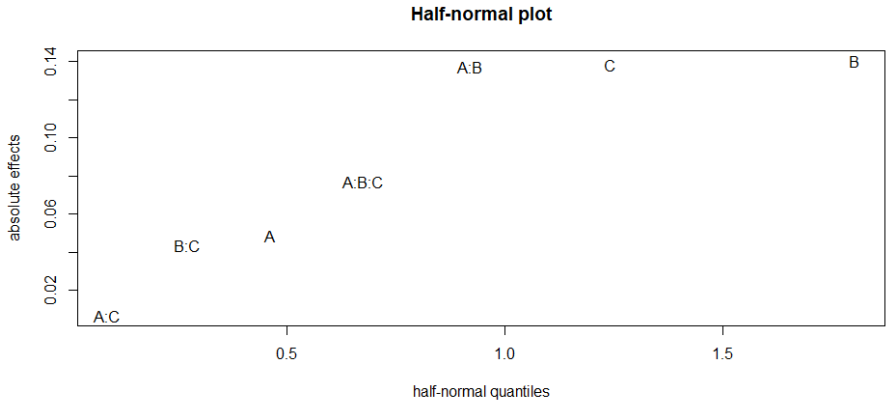

Figure 2.6: *half-normal plot*

From the half-normal plot and Lenth's method, main effect B,C and interaction effect AB are significant factors. Therefore, the model has three terms. The R^2 value for this model is 0.849 and the adjusted R^2 value for this model is 0.7358. The P value for the F-test is 0.04052 and the P value for AB,B and C is 0.05335, 0.05028 and 0.05241. The explicit expression of the model is

$$CR_1 = 0.12655 - 0.06863AB + 0.07009B + 0.06907C$$

Similarly, for funnel 2, the half-normal plot and $|t_{pse}|$ from Lenth's method for funnel 2

Effect	$\lvert t_{PSE} \rvert$
A	0.1558
C	0.6863
AC	0.6667

Table 2.9: $|t_{PSE}|$ *Values for multivariate testing*

are given in figure 2.7 and table 2.9

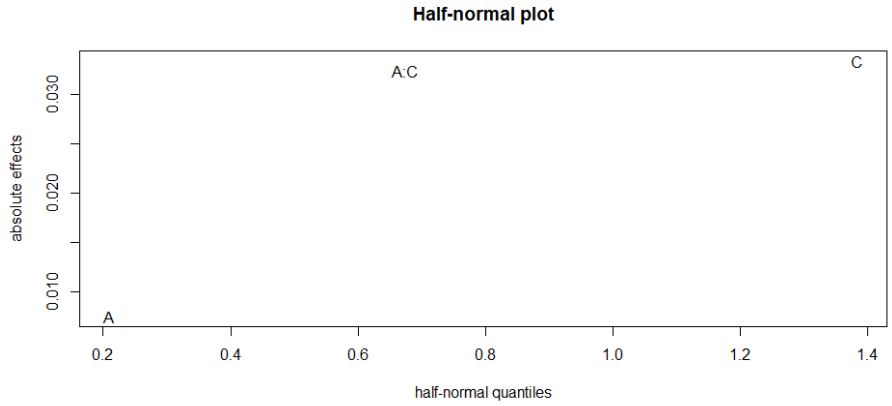

Figure 2.7: *half-normal plot*

We choose main effect C and interaction effect AC as significant factors. The R^2 value for this model is 0.9058 and the adjusted R^2 value for this model is 0.8681. The P value for the F-test is 0.002724 and the P value for C and AC is 0.0042 and 0.00475. The explicit expression of the model is

$$CR_2 = 0.091215 - 0.016675C - 0.016198AC$$

The last step of traditional funnel testing is to combine the expression from the two conversion funnel and find the optimal settings of each factor. Since factor C belong to two different conversion funnels, additional copy of C is created as C'. The final expression omit

intercept term for the conversion system is therefore

$$CR_{total} = -0.06863AB + 0.07009B + 0.06907C - 0.016675C' - 0.016198AC'$$

We now analyze the system using CME-based funnel testing. For conversion funnel 1, variable A, $B \mid A+$, $B \mid A-$, $C \mid B+$, $C \mid B-$ are included as our candidate variables. Stepwise regression with forward selection is performed and variable $B \mid A-$, $C \mid B+$ and $C \mid B-$ are added to the model. The final model only contains three terms. The R^2 value for this model is 0.8773 and the adjusted R^2 value for this model is 0.7853, **which are higher compare to traditional funnel testing**. The P value for the F-test is 0.02705 and the P value for $B \mid A-$,$C \mid B+$ and $C \mid B-$ is 0.01263, 0.04778 and 0.1928. The explicit expression of the model is

$$CR_1 = 0.12336 + 0.13226B|A^- + 0.07967C|B^+ + 0.04702C|B^-$$

Similarly, the CME covariates for funnel 2 are A, $C|A-$ and $C|A+$. After stepwise regression, only $C|A+$ in selected to include in the model. The R^2 value for this model is 0.93 and the adjusted R^2 value for this model is 0.9183, The P value for the F-test is 0.00011 and the P value for $C|A+$ is 0.00011. The explicit expression of the model is

$$CR_2 = 0.0925 - 0.04745C|A+$$

. Combine the expression with the expression from funnel 1 analysis, the final modeling

result is

$$CR_{total} = 0.13226B|A^- + 0.07967C|B^+ + 0.04702C|B^- - 0.04745C|A^+$$

.

Remark: from the probability distribution of funnel 2, $t_1 = 0.5 - 0.1A$ and $t_{ac} = 0.38 - 0.1C + 0.05A - 0.1AC$. It is clear that C should be set at negative level when A is set at positive level. When A is set as negative level, C's level setting no longer matter. The second model from CME $CR_2 = 0.0925 - 0.04745C|A+$ captures the true level for C based on A and give the experiment observer a clear picture of how the between page interaction works. Compare to traditional funnel testing $CR_2 = 0.091215 - 0.016675C - 0.016198AC$, which did not give a clear picture of how C's level setting depends on A's level setting. Although the funnel testing gives the correct factor setting in this case for funnel 2(A+ and C-), after combine with funnel one A's level will be set to negative and C's level will be wrongly set to negative, where in the truth it should not matter. The CME-based funnel testing achieves better modeling and optimization result and enjoys simplicity compare to traditional funnel testing, since CME-based funnel testing does not have to creating additional copies for variables belong to multiple conversion funnels.

2.5.3 Example 3: System with multiple conversion funnels contains factors with 3 level or above

Now considering a website system consisting of six pages as denoted by figure 2.8.

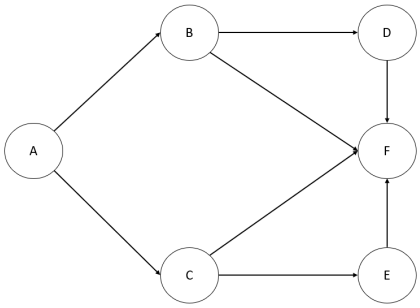

Figure 2.8: *Complex conversion system*

We assume A is the factor on homepage where it has 3 options that needs to be compared. For example, factor A may represent the choice of the header, which has three candidates versions. B and C are the factors on two different product listing pages, where multiple products are listed for customer to compare based on their categories. B has three levels and C has two levels. D and E are the factors on subsequent product information pages, in where detailed product information for specific product is presented, which has 2 levels for each of them. F is the factor on checkout page where it has 3 levels.

A mixed $OA(36, 2^3 3^3)$ design is used for this experiment. The design matrix is shown in table 2.10.

The next step is to identify all the conversion funnels from homepage to the conversion page. From figure 2.8, there are four conversion funnels. We call $A \to B \to D \to F$ as funnel I; $A \to B \to F$ as funnel II; $A \to C \to F$ as funnel III and $A \to C \to E \to F$ as funnel IV. In this example we will show how to analyze funnel 1 using the framework proposed in section 3, the analysis strategies can be similarly applied to rest of the conversion funnels.

For funnel 1, $A \to B \to D \to F$, there are 3 variables with level 3. In this example we

A	B	C	D	E	F
-1	-1	-1	-1	+1	0
-1	-1	+1	-1	-1	0
-1	-1	-1	+1	-1	0
-1	-1	+1	+1	+1	0
0	-1	-1	-1	+1	+1
0	-1	+1	-1	-1	+1
0	-1	-1	+1	-1	+1
0	-1	+1	+1	+1	+1
1	-1	-1	-1	+1	-1
1	-1	+1	-1	-1	-1
1	-1	-1	+1	-1	-1
1	-1	+1	+1	+1	-1
-1	0	-1	-1	+1	+1
-1	0	+1	-1	-1	+1
-1	0	-1	+1	-1	+1
-1	0	+1	+1	+1	+1
0	0	-1	-1	+1	-1
0	0	+1	-1	-1	-1
0	0	-1	+1	-1	-1
0	0	+1	+1	+1	-1
1	0	-1	-1	+1	0
1	0	+1	-1	-1	0
1	0	-1	+1	-1	0
1	0	+1	+1	+1	0
-1	+1	-1	-1	+1	-1
-1	+1	+1	-1	-1	-1
-1	+1	-1	+1	-1	-1
-1	+1	+1	+1	+1	-1
0	+1	-1	-1	+1	0
0	+1	+1	-1	-1	0
0	+1	-1	+1	-1	0
0	+1	+1	+1	+1	0
1	+1	-1	-1	+1	+1
1	+1	+1	-1	-1	+1
1	+1	-1	+1	-1	+1
1	+1	+1	+1	+1	+1

Table 2.10: *Design matrix for complex example*

use

$$D_{01} = \begin{cases} -1 & \text{for level 0} \\ 1 & \text{for level 1} \\ 0 & \text{for level 2} \end{cases}$$

$$D_{02} = \begin{cases} -1 & \text{for level 0} \\ 0 & \text{for level 1} \\ 1 & \text{for level 2} \end{cases}$$

to represent the two main effect contrast for the three level factor A,B and F. The candidate variables set then contains

$$B_{01}|A_0, B_{01}|A_1, B_{01}|A_2, B_{02}|A_0, B_{02}|A_1, B_{02}|A_2, D|B_0, D|B_1$$

$$D|B_2, F_{01}|B_0, F_{01}|B_1, F_{01}|B_2, F_{02}|B_0, F_{02}|B_1, F_{02}|B_2$$

Stepwise regression then performed and the final model is expressed as

$$CR_1 = 0.0055391 + 0.0717B_{02}|A_0 + 0.03552D|B_1 + 0.0339D|B_2 + 0.0299B_{02}|A_1 + 0.018132D|B_0$$

The analysis strategy is applied to all the conversion funnels. $CR_2 = 0.31884 + 0.2875B_{02}|A_0 - 0.20399F_{02}|B_2 + 0.13892B_{02}|A_1$, $CR_3 = 0.16239 + 0.16166C|A_0 + 0.14305F_{02}|C_2 + 0.11246C|A_1 + 0.06067C|A_2 + 0.03653F_{02}|C_1$, $CR_4 = 0.05586 + 0.06520C|A_0 + 0.03621F_{02}|E_1 + 0.03345F_{02}|E_2 + 0.02265C|A_1$. Then the conversion rate models are combined to form the

Functions for decision probabilities for setting 1
$t_1 = 0.5 + a * A$
$t_{ab} = 0.5 + b * B + ab * AB$
$t_{bc} = 0.38 + c * C + bc * BC$

Table 2.11: *Probability distribution*

Functions for decision probabilities for setting 2
$t_1 = 0.4 + a * A$
$t_{ab} = 0.5 + b * B + ab * AB + a2 * A$
$t_{bc} = 0.38 + c * C + bc * BC + b2 * B$

Table 2.12: *Probability distribution*

final model CR_t by weighted average or each model can be optimized to find the optimal setting for each individual conversion funnel.

2.6 Numerical studies

From the above examples, We see CME-based funnel testing enjoys better model sparsity, higher adjusted R^2, and smaller p-values on the coefficients. To study its optimization accuracy, numerical simulations are conducted. Only single funnel optimization is considered here because multiple funnels' optimization accuracy directly depend on the optimization performance for each individual conversion funnel. We consider two general settings, setting 1 is defined as table 2.11. where a,b,ab,c,bc are random generating coefficients satisfying the constrain $0 \leq t_i \leq 1$. Each A,B,C has two levels and 2^3 full-factorial design is used.

Another setting is also considered where the probability function is showed in table 2.12. In this simulation, A,C has two levels and B is assumed to have 3 levels. An $OA(12, 2^2 3)$ design is used to conduct the experiment. For each setting, 500 simulation

	accurate percentage	relative loss
setting 1	91.2%	8.21%
setting 2	89.4%	6.32%

Table 2.13: *Accuracy measures for CFO*

are performed and the percentage of CFO find the best factorial settings is recorded. If the CFO misses the best factorial setting, the relative loss is record based on the formula $(maximum\ \ conversion\ \ rate - CME\ \ seleted\ \ conversion\ \ rate)/maximum\ \ conversion\ \ ra$ The result is presented in table 2.13

We can see the CME-based funnel testing achieves satisfying result both in terms of accuracy percentage and relative loss.

2.7 Conclusion

In this paper, a novel CRO framework based on CME, called CFO, is presented. The proposed framework enjoys good interpretability, simplicity and optimization accuracy. Popular framework like A/B testing enjoys simplicity and interpretability but lacks optimization accuracy. The model-free multivariate testing enjoys optimization accuracy and simplicity but lacks interpretability. We demonstrated the CME-based funnel testing achieves better modeling result compare to model-based multivariate testing and traditional funnel testing. Furthermore, the CFO can be done in an automatic way, while the traditional funnel testing requires hand-on manipulation.

Looking forward, there are many research directions to pursue next. Conditional main effect based on levels choices which are large than three should be studied. The Markovian properties maybe relaxed in real world application, but how to handle the exponentially

growing variable size remains unclear at this point. Finally, an R-package based on CFO should be made available to the general public so the practitioner can take advantage of our method and gives valuable feedback.

	Interpretability	Simplicity	Accuracy
A/B testing	✓✓	✓✓	X
model-free MVT	X	✓✓	✓✓
traditional funnel testing	✓	X	✓
CFO	✓✓	✓	✓

Table 2.14: *Performance summarization*

CHAPTER 3

SARAN: SEQUENTIAL ADAPTIVE RADIAL BASIS FUNCTION NETWORK BASED EMULATOR FOR NON-STATIONARY, LARGE SCALE EXPERIMENTS

3.1 Abstract

The Gaussian process is a standard tool for building emulators for computer experiments. However, due to its lack of ability to model large-scale and non-stationary data, Gaussian process is greatly limited in practice. We provide a new approach to approximate emulation of large computer experiments. By taking advantage of the learning ability and strong tolerance to input noise of radial basis function, we derive a sequential learning scheme that dynamically optimize the basis function's location, scale and coefficient. L-1 penalty is utilized to ensure our emulator's simplicity. We applied our method to study solar irradiance computer model based on half million physical measurements data. We demonstrate that the proposed model enjoy marked advantage over existing emulation tools in both emulation accuracy and data capability in terms of non-nationality and sample size.

3.2 Introduction

Computer experiment is an experiment used to study a computer simulation, which is an implementation of complex mathematical models using computer codes. Computer simulations are used to study system of interest for which physical experimentations are infeasible or limited. However, computer simulations are usually developed based on

simplifying assumptions of the physical system. In some cases, the physical system is so complicated that it cannot be described by mathematical models. Therefore, the predictions based on computer models can often go wrong when the assumptions are violated or the computer model is not adequate to describe the actual physical system. Thus, to make the predictions meaningful and closer to reality, the model calibration process need to be performed to correct the bias from computer model.

In a fundamental work, Kennedy and O'Hagan [27] proposed a Gaussian process-based Bayesian framework for doing model calibration. [28] [27]. Gaussian process [29] modeling build *emulator* [30],[29] that leverages known properties of the underlying response surface to produce both predictions and uncertainty quantification for the prediction. Gaussian process emulation is mathematically simple and enables statistical uncertainty quantification via confidence intervals. However, it is known [31] that Gaussian process is not only hard to compute when number of sample goes large, but also produces highly unstable result when input data is non-stationary. Those bottlenecks makes Gaussian process unattractive for large-scale computer experiments.

In some applications, the input data from both computer experiments and field can be non-space filling, non-stationary, and has enormous volume. This article describe a new radial basis function network [32] approach to build emulator for large-scale computer experiments. We called the new method **SARAN**: **S**equential **A**daptive **RA**dial basis function **N**etwork. **SARAN** automatically update it's scale and location to utilize the structure in our dataset by putting few wide basis functions to place where data distribution are relatively stationary and putting many narrow basis functions to place where data distribution are non-stationary.

The remainder of this article is organized as follows: The motivating application on

predicting solar irradiance as function of location and time is described in section 2. The novel sequential adaptive radial basis function network is formalized and proposed in section 3 where the algorithm for fitting radial-basis function as well as a newton-like fast optimization algorithm is provided. The results of applying the proposed method to both the simulated and the real data is also included in section 3. In Section 4, an ensemble estimator is proposed to combine the biased estimation from computer experiment with the unbiased estimation from measured data. At last, summary and future works are given in Section 5.

3.3 Motivating Application: Solar irradiance prediction

A key component in power balancing and determining the viability of potential sites for harvesting solar power is predicting solar irradiances, or the power per unit area produced by electromagnetic radiation, as a function of the geo-locations. Previous work has been focused on short-term forecasting for solar irradiance. [33] Uses machine learning and time series model to forecast hourly solar irradiance. [34] provides framework for day-ahread solar forecasting. [35],[36] has used meteorological variables like temperature, cloud cover, and wind speed to predicting solar power using neural networks and some other machine learning algorithms. We are working with daily solar irradiance data from three sources co-located at 1535 weather stations distributed throughout the continental of United States, from September 18, 2014 to April 15, 2016. Those stations were taken from the selected sites in the Remote Automated Weather Stations (RAWS) network[37]. The data is collected in every 15 minutes. Please note that the weather stations are not uniformly distributed in the United States. In particular, many promising locations for solar farms are sparsely covered. A visualization of the weather station location is provided as black dots in Figure

3.1. The three data sources include:

1. Global horizontal irradiance (GHI) measurements at a specturm of fixed geographic locations through time;

2. Output from North American Mesoscale Forecast system (NAM; http://www.ncdc.noaa.gov/node/54),

3. Output from the Short Range Ensemble Forecast (SREF; http://goo.gl/vsE8Yi), as well as clear sky solar irradiance, zenith, and azimuth values.

The first data source (GHI) is provided by equipment housed at the weather stations, which is the actual measured physical data. The second (NAM) and third (SREF) data are simulation-based irradiance sources coming from computer models.

There are several characteristics of the solar irradiance measurement data that makes it hard to model. First is the size of the dataset, there are more than 18 million rows in the dataset and over 5GB in storage space. It is computationally unfeasible for procedure like Gaussian process. Second is the non-space-filling-design properties of the measured data. The solar sites are more crowded in places like West coast of the United states compared to the middle east region. The third challenge is the non-stationarity in both space and time. Intuitively, places that are close to each other should have less differences compare to places that far away. However, a myriad of geological features, such as large mountains, lakes, etc, can significantly change the angle of the sun and then change the solar intensity. Figure 3.1 shows the quantiles over measured data indicated by color (blue indicate low value, green indicate medium-low value, yellow indicate medium-high value and red indicate high values for the measured data).

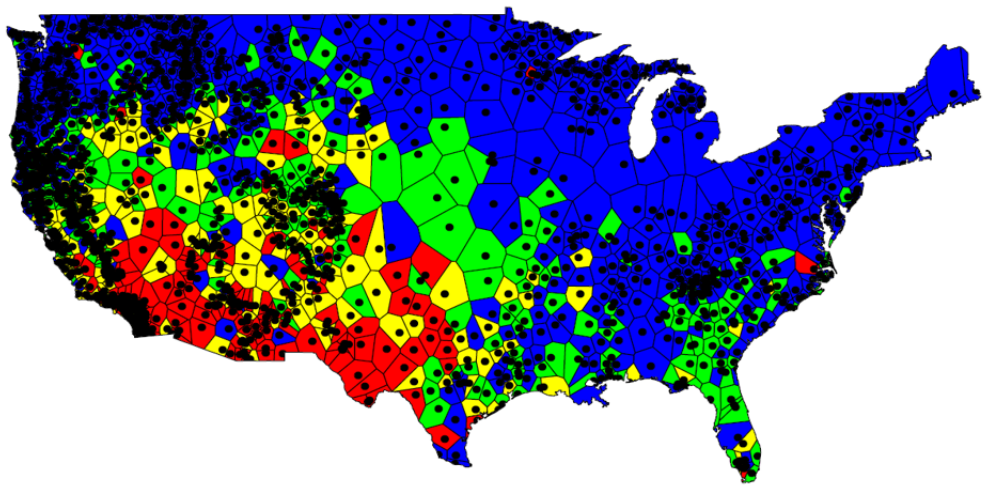

Figure 3.1: *Yearly average of measured data.*

Unfortunately the model predictions from NAM and SREF, which are based on weather models, are poorly calibrated. They either match one another nor match the expectation of the measured solar irradiances at the fixed geographic locations. See as an example figure 1, which shows 3 days of data including actual measurements, and predictions from the models at 3 sample locations. 3.2 and the yearly average quantile map from yearly average of NAM and SREF. 3.3 3.4

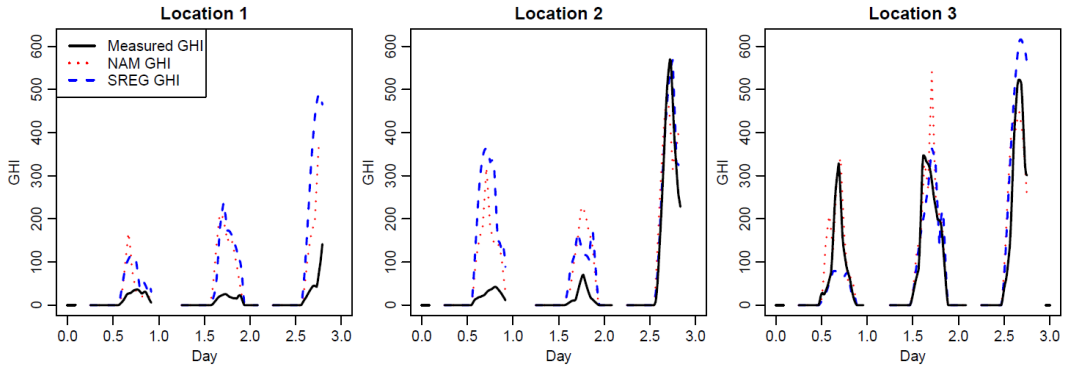

Figure 3.2: *Three days of measurement data, NAM and SREF models at 3 sample locations.*

Our interest lies in predicting the expectation of measured solar irradiances over the whole map of United States by combining the information provided by data from actual

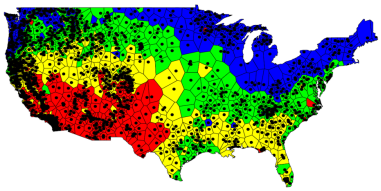

Figure 3.3: *Yearly average from NAM forecast*

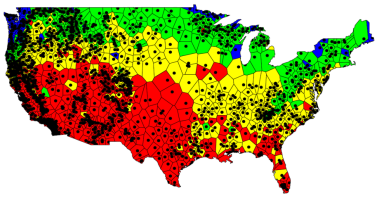

Figure 3.4: *Yearly average from SREF forecast*

physical measurements and data from the two weather models. Intuitively, the measured data is unbiased, but has a lot of noise; The data from weather model are less noisy but has large bias as demonstrated in Figure 3.2. We propose the following calibration model:

$$y_1(s_1, s_2) = w_1 f_1(s_1, s_2) + w_2 f_2(s_1, s_2) + w_3 f(s_1, s_2) + \epsilon$$

where s_1, s_2 denotes the longitude and latitude or the geographic measurement, f_1, f_2 and f represents the NAM, SREF and Measured data for solar irradiance, and ϵ_t represents the stochastic error. Conditional on the estimates for w_1, w_2, w_3, e.g. via least squares, the estimation of f_1, f_2, f is challenged by the large inherent sample sizes and non-stationarity.

3.4 SARAN: Sequential Adaptive radial basis function network

In this section, we present our novel statistical estimation method SARAN. A radial basis function (RBF) is a real-valued function ϕ whose value depends on the distance from some point c, called a *center*, and the parameter to adjust the value of the distance θ, called *scale*, so that $\phi(x, c, \theta) = \phi(\theta_k(x - c_k))$. Radial basis functions are typically used to build up functional approximation of the form $\hat{y}(x) = \sum_{k=1}^{K} \beta_k \phi(\theta_k(x - c_k))$ where the approximating function $\hat{y}(x)$ is represented as a sum of K radial basis functions, each associated with a different center c_k and scale parameter θ_k. The basis functions are weighted

by coefficient β_k. The weight β_k are usually estimated by least square method because the approximating function is linear in the weights β_k.

Traditional radial basis function requires user to input the center and scale parameter as well as the number of basis functions. Once the number of basis function is fixed, only the coefficients will be driven by the data. Intuitively, in regions where the function is slowly varying, we need relatively few, wider basis functions. In regions where the function is quickly varying, we need relatively more, narrower basis functions. We would like data to drive the size, location, and number of basis functions. In the absence of information, we would like to encourage relatively few functions, so that information is tied together smoothly over data gaps.

Write $\hat{y}(x) = \sum_{k=1}^{K} \beta_k \phi(\theta_k(x - c_k))$ for some bell-shaped kernel ϕ and consider squared error loss $||y_i - \hat{y}(x_i)||^2$. We take a convex regularization perspective as a computationally efficient means to encourage as much simplicity in the model as the data will allow. In particular, we use L_1 penalty on β to encourage sparsity on the number of basis function. The overall objective function is then

$$Q = min_{\{}c_k, \theta_k, \beta_k\} \frac{1}{N} \sum_{i=1}^{N} \left(||y_i - \sum_{k=1}^{K} \beta_k \phi(\theta_k(x - c_k))||^2 \right) + \lambda_1 ||\beta||_1 \tag{3.1}$$

Ideally, we would like to optimize the sum-of-squares with respect to all the parameters. Unfortunately, the criterion is non-convex with multiple local minima. There does not exist any algorithm to estimate c, θ, β at the same time. We propose to use the following optimization method to estimate the parameter sequentially

3.4.1 Sequential approximation method

Due to the non-convexity of Eq 3.1. A sequential algorithm is proposed to estimate $\{c_k, \theta_k\}$ separately from the β_k. Given the former, the estimation of the latter is a least squares problem with l_1 penalty, where algorithm like *lars* [38] can solve it efficiently. However, estimate $\{c_k, \theta_k\}$ is still a very challenge problem in big data setting. We consider *linearizations* of $\hat{y}(x)$ (as a function of parameters) to allow efficient computation of parameter updates within iterations. Let ϑ denote a vector of parameters for updating and write

$$\begin{aligned}
\hat{y}_{\vartheta_{new}}(X) &\approx \hat{y}_{\vartheta_{old}}(X) + \frac{\partial \hat{y}_{\theta_{old}}(X)}{\partial \vartheta'_{old}}(\vartheta_{new} - \vartheta_{old}) \\
&= \hat{y}_{\vartheta_{old}}(X) - \frac{\partial \hat{y}_{\theta_{old}}(X)}{\partial \vartheta'_{old}}\vartheta_{old} + \frac{\partial \hat{y}_{\theta_{old}}(X)}{\partial \vartheta'_{old}}\vartheta_{new} \\
Q &= ||y - \hat{y}_{\vartheta_{new}}(X)||^2 \\
&\approx ||y - \hat{y}_{\vartheta_{old}}(X) - \frac{\partial \hat{y}_{\theta_{old}}(X)}{\partial \vartheta'_{old}}\vartheta_{old} + \frac{\partial \hat{y}_{\theta_{old}}(X)}{\partial \vartheta'_{old}}\vartheta_{new}||^2 \\
&= ||y^* - x^*\beta^*||^2
\end{aligned}$$

Then, for basis function widths (θ) and locations (c), one has closed form updates

$$\theta_{new} = (X'_* X_*)^{-1} X'_* y_*, \quad X_* = \frac{\partial \hat{y}_{\theta_{old}}(X)}{\partial \theta'_{old}}, \quad y_* = y - \hat{y}_{\theta_{old}} + \frac{\partial \hat{y}_{\theta_{old}}(X)}{\partial \theta'_{old}}\theta_{old}$$

and

$$c_{new} = (X'_* X_*)^{-1} X'_* y_*, \quad X_* = \frac{\partial \hat{y}_{c_{old}}(X)}{\partial \theta'_{old}}, \quad y_* = y - \hat{y}_{c_{old}} + \frac{\partial \hat{y}_{c_{old}}(X)}{\partial c'_{old}} c_{old}$$

while for basis function coefficients (β), we have an L_1 regularized linear regression(lasso) problem [39]. Computationally efficient algorithm (LARs) [38] exist for solving the above problem in the big data context.

See detailed algorithm as Algorithm 3.4.1.

A few details are notable. First, the k^{th} column of $\frac{\partial \hat{y}_{\beta_{\text{old}}}(X)}{\partial \beta'_{\text{old}}}$ is $\phi(\theta_k(X - c_k))$. Derivatives with respect to θ and c can commonly be written in closed form, but it is simpler (and more stable) to use numeric differentiation. Selection of λ_1 could be via k-fold cross-validation, which can be parallelized easily in a moderately big data setting (or estimated via a separate testing set in a *very* big data setting). Finally, parameters need to be initialized (and directed) with care to ensure convergence to a high-quality solution. Our current prototype code initializes with $K = n/10$ basis functions centered at $n/10$ randomly selected input values, each with fixed width. After each update, if the error does not reduce enough, additional candidate basis functions centered near regions with larger errors are introduced into the candidate set.

Algorithm 3: BasisMovement Newton Version

Initialization: start with roughly 10 percent basis functions, the initial location C is the data points and initial scale parameter θ are 1. Took out roughly 20 percent data as testing and the rest are used as training data.

while *not converge to the true function* **do**

find the coefficient

$\hat{u}(x) = \sum_{j=1}^{n} \beta_j \phi_{\theta_j}(C_j, x)$ where n is the number of basis function, where $\phi_{\theta_j}(C_j, x) = exp\{-\theta_j(C_j - x)^2\}$ Then we have close form solution for the β which is $\beta = (\phi^T\phi)^{-1}\phi^T y$

find all the locations and then all the scale parameter through Newton optimization method

Using

$$\hat{u}_{\theta_{new}}(x) \approx \hat{u}_{\theta_{old}}(x) + \frac{\partial \hat{u}_{\theta_{old}}(x)}{\partial \theta_{old}}(\theta_{new} - \theta_{old})$$

to find C

$$Q = ||y - \hat{u}_{C_{old}}(x) + \frac{\partial \hat{u}_{C_{old}}(x)}{\partial C_{old}}(C_{new} - C_{old})||^2$$

$$= ||y^* - x^*\beta^*||^2$$

where $X^* = \frac{\partial \hat{u}_{C_{old}}(x)}{\partial C_{old}}$ (Jacobian matrix) $y^* = y - \hat{u}_{C_{old}}(x) + \frac{\partial \hat{u}_{C_{old}}(x)}{\partial C_{old}} C_{old}$ and $\beta^* = C_{new}$

to find θ

$$Q = ||y - \hat{u}_{\theta_{old}}(x) + \frac{\partial \hat{u}_{\theta_{old}}(x)}{\partial \theta_{old}}(\theta_{new} - \theta_{old})||^2$$

$$= ||y^* - x^*\beta^*||^2$$

where$X^* = \frac{\partial \hat{u}_{\theta_{old}}(x)}{\partial \theta_{old}}$ $y^* = y - \hat{u}_{\theta_{old}}(x) + \frac{\partial \hat{u}_{\theta_{old}}(x)}{\partial \theta_{old}} \theta_{old}$ and $\beta^* = \theta_{new} = (X^{*T}X^*)^{-1}X^{*T}y^*$

use Lasso penalization to select basis function

Use lars package to find the entire lasso path, use 10 fold cross-validation to choose the best fraction of coefficient and then use that fraction to choose the best turning parameter.

Throw in basis function to ensure convergence

if $(mseError_{t-1} - mseError_t < \epsilon)$

Add basis function with center D and theta draw from multinomial distribution with probability density based on training errors. The number of basis function N is based on the number of basis function lasso in the previous step decide to throw away, define as DN. Currently I choose $N = 0.5 * DN$.

end

3.4.2 Results on simulated example

The current goal is to emulate function

$$y = exp((x + 0.5)^2)sin(exp((x + 0.5)^2))$$

Which looks like Figure 3.5.

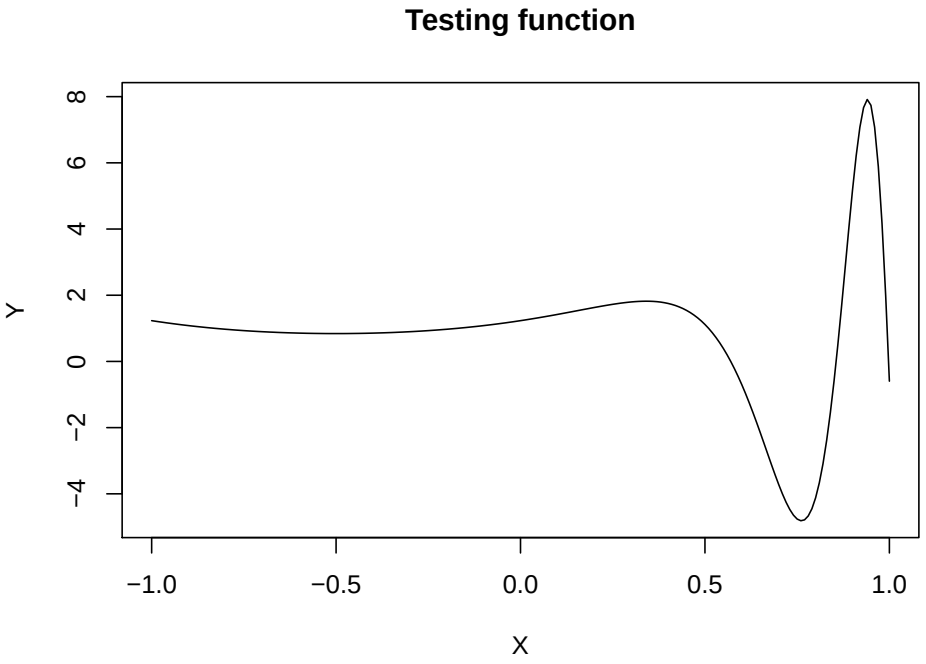

Figure 3.5: *Sample function to emulate*

To demonstrate the superiority of our method, we first fit this data using *loess* function with *stats* package in R [40] and *gam* package in R [41] with spline as basis function. We can see from figure 3.6 and 3.7 that the result is not desirable.

We fit the data using our algorithm, Figure 3.8 shows the fitting algorithm in iterations 1, 3, 5, and 50. Initially, there are far too many basis functions and the fit for the ill-conditioned

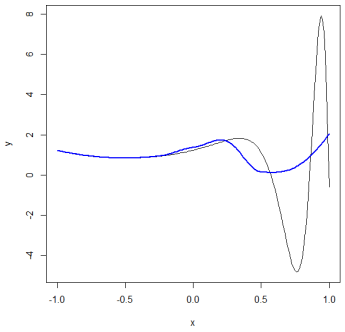

Figure 3.6: *blue indicate loess function fitting result*

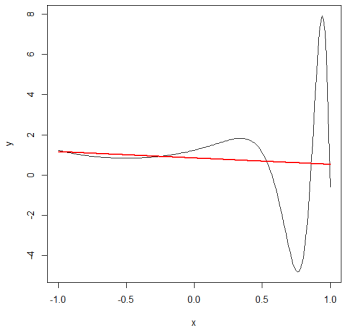

Figure 3.7: *red indicate spline function fitting result*

problem is very poor. By iteration 5, the fit is quite good, and most of the basis functions have been eliminated, with those near slowly varying regions wider and those near more quickly varying regions narrower. By iteration 50, the model is a near interpolator and only 7 basis functions remain.

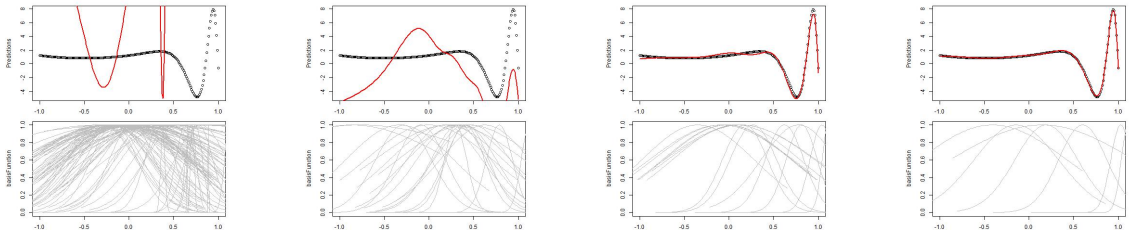

Figure 3.8: *Panels from left to right show iterations 1, 3, 5, and 50 of radial basis function network fitting algorithm. Upper panels show data (circles) and predictions (red curve). Lower panels show selected basis functions.*

3.4.3 Fitting result on time-aggregated solar irradiance data

The measurements data we have are collected in every 15 minutes, but we focus our study on the yearly average solar irradiance data for a variety of reasons: the first reason is the purpose of the solar irradiance study. One of the main purpose for forecasting solar irradiance is to identify promising locations for economical solar farming. Correctly identify the locations that can accumulate most sun-light on a yearly basis can help the government

or companies find the best location for solar power collection. Another reason is the size and the missingness of the data. About 17 % of locations are missing more than 5% of daily observations, and no location is fully observed, which makes the time-series model very challenging. Therefore for each of the 1535 spatial locations, we average the irradiance values for both the measured data and the NAM and SREF computer models. We leave the complete spatio-temporal model of solar irradiance based on the original data resolution as future work.

We use 10-fold cross-validation [42] to measure the performance of our model: iteratively hold out about one tenth of the particular spatial locations, train on the remaining data, and make a prediction for the held-out locations. We measure accuracy of each prediction in terms of *root-mean-squared error(RMSE)* based on the average of 10 cross-validation iterations, as denote in Eq 3.2, where n is the number of samples in the held-out data for each cross-validation iteration.

$$\frac{1}{T}\sum_{t=1}^{T=10}\sqrt{\frac{1}{n}\sum_{i\in testing(t)}(y_i-\hat{y_i})} \tag{3.2}$$

Where $T = 10$ denotes the number of iterations of cross-validation, *testing(t)* denotes the set of testing indices for cross-validation iteration *t*, and y_i is the testing data output and $\hat{y_i}$ is the predicted value on testing data input.

All the computation are carried out in **R**, We compared SARAN with the following models to serve as baselines

1. ordinary GP models [43] implemented in the **mlegp** package [44].

2. local GP process [45] implemented in the **laGP** package [46].

3. general additive model with smoothing spline [47] in the **gam** package [41].

4. local linear regression [48] in the **stats** package [40].

Please note that beating all the baseline by large margin is not the main focus of this section since most model's behavior can be largely improved by tuning. They are here to serve the purpose of bench-marking to make sure our method does not perform poorly. The result is summarized in table 3.1.

target	SARAN	mlegp	laGP	gam	loess	mean & sd	NAM	SREF
measured	19.04	34.99	20.11	19.56	20.69	168.5 & 37.88	32.79	44.96
NAM	9.03	27.34	8.52	10.98	13.43	190.2 & 27.35		
SREF	8.56	23.86	9.57	11.23	10.87	205.9 & 26.26		

Table 3.1: *10-fold cross-validation average RMSE. The "mean & sd" column is the summary statistics about measured data, NAM data and SREF data with left:mean, right:standard deviation; the last two columns is the RMSE value from using NAM and SREF data directly to predict measured data*

From table 3.1, **SARAN** performs very well compared with benchmarks, beating all the baseline methods 2 out of 3 times on measured data and SREF data. On NAM data, **SARAN** has a slightly higher RMSE compared to **laGP** but still better than the rest of the baseline methods. We plot the measured data average alongside with the cross-validation predicted average in Figure 3.9 and Figure 3.10. We can see the predicted value captures most of the non-stationarity and spatial trend of the measured data.

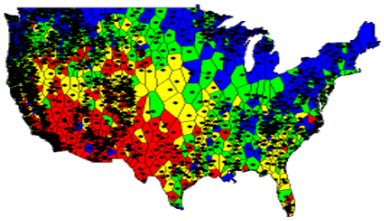

Figure 3.9: *Training measured average*

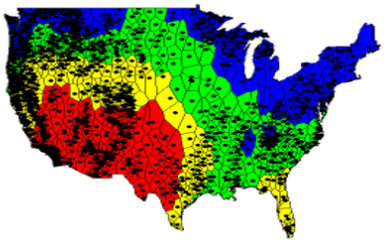

Figure 3.10: *CV predicted average*

3.5 Combining estimation from physical and computer experiments

Intuitively, measured field data has less bias but large variance, and the computer model output has smaller variance but larger bias compared to field data. We would like to combine the information from computer model output, i.e. NAM and SREF with the measured data to get a better model. The first thing comes into our mind is following the literature in computer model *calibration* [27] [28]. Computer model calibration invloves model the discrepancy between the computer simulated data and the measured data. Kennedy and O'Hagan [27] proposed a Bayesian framework which can "combine" these two data sources. However, the Kennedy and O'Hagan framework is not suitable for the synthesis of the three solar irradiance data for a variety of reasons. First, the goal is different, we do not have tuning parameters to adjust, our main goal here is to simultaneously combine the existing data sources to get a better predictor. Second, the data from all three sources are all too big for Gaussian process modeling. Besides, the stationary assumption from Gaussian Process will be unrealistic given the non-stationarity in our data. Although many methods of relaxing stationary assumption exists in the literature [49] [50], few offers software to implement it.

The usual combined estimator assume both estimator are unbiased and average with weights which are inversely proportional to the variance. If all the estimators are indepen-

dent, the *inverse-variance weighting* method is known to be optimal [51]. In our problem set up, we have one estimate from physical experiment which is believed to be unbiased, and we have another two estimates from computer simulation, which are believed to be biased. De-bias the computer model output from physical measured data will result in dependency between the estimators, which makes the *inverse-variance weighting* independency assumption not hold.

Assume $X \sim \mathcal{N}(\theta, \sigma^2)$ as the unbiased estimator, $Y_1 \sim \mathcal{N}(\theta + \xi_1, \tau_1)$ and $Y_1 \sim \mathcal{N}(\theta + \xi_2, \tau_2)$ for the biased estimator with ξ_1 and ξ_2 as their respective bias. Another method to combine biased and unbiased estimator is the James-stein shrinkage type estimator [52]. However, the shrinkage estimator needs accurate estimation about the variance term σ^2, τ_1, tau_2 and bias terms ξ_1, ξ_2. This is not realistic given the non-stationarity in our data sources. Instead, we borrow ideas from machine learning literature to create a *ensemble* estimator from the prediction of measured data, NAM and SREF.

Intuitively, the computer model captures some of the non-stationarity of measured data, so fitting their discrepancy would be easier comparing to fitting measured data alone. For each computer model (NAM and SREF), we estimate their discrepancy from the true mean function, This is done by using SRBF to fit measured data minus computer model data $X - Y_i$; After we have the two discrepancy functions, we construct two estimates from the true mean function at a new location and time by evaluation the relevant computer model and then adding the relevant discrepancy function. Out third estimator will be from building the emulator for the measured data directly using SARAN.

We formulate the problem into the following optimization problem, let $\hat{\mu}_m(x_i)$, $m = 1, 2, 3$ denotes the predictor build from measured data, NAM and SREF using **SARAN**.

Let w_m denotes the weight for the corresponding predictors. The we can write down the optimization problem as follow:

$$\begin{aligned} \min_{w_m} \quad & \sum_{t=1}^{T} \sum_{i \in testing(t)} (y_i - \sum_{m=1}^{3} w_m \hat{\mu}_m^t(x_i))^2 \\ \text{s.t.} \quad & \sum_{m=1}^{3} w_m = 1 \\ & w_m \geq 0, m = 1, 2, 3 \end{aligned} \tag{3.3}$$

Where T denote the number of iterations of cross-validation, *testing(t)* denote the testing indices for cross-validation iteration t, and $\hat{\mu}_m^t$ denotes the prediction model based on the data not in *testing(t)*. We randomly sampled one tenth of the data as held-out testing set and solve the optimization problem on the training set. The RMSE error on the testing set is 18.23, compared to 19.56 when using **SARAN** function directly on the measured data. We plot the prediction on the whole dataset using ensemble method in Figure 3.11

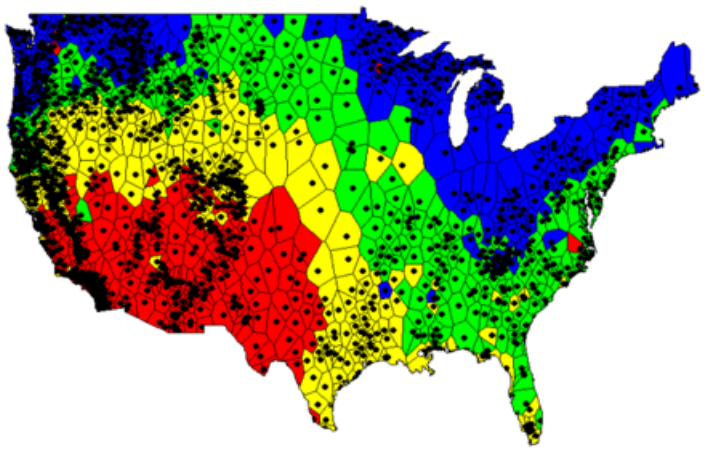

Figure 3.11: *prediction using the entire dataset*

we can see compared with Figure 3.12 and 3.13, the prediction by ensemble model captures more non-stationarity (in the south-west region) than fitting measured data alone

using **SARAN**.

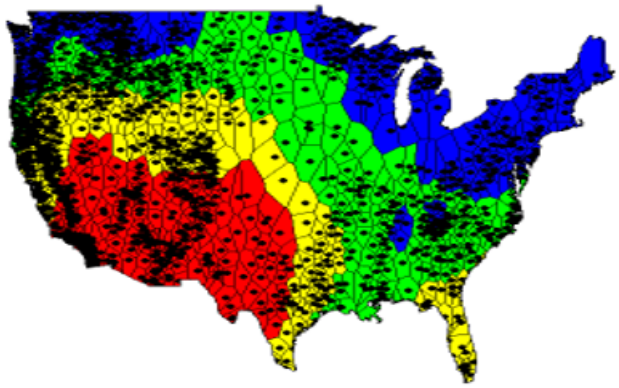

Figure 3.12: *CV prediction using measured data alone*

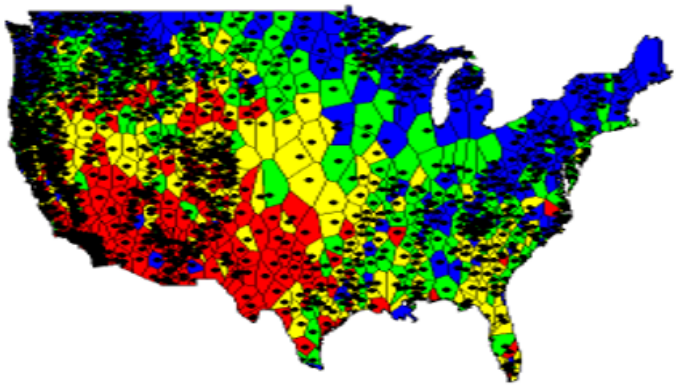

Figure 3.13: *The ground truth measured data*

3.6 Summary and future works

While large-scale and non-stationary problems have become typical in the modern "big data" era, Gaussian process models are often impractical due to the memory issues and numerical instability. In this paper, we proposed a novel method, called **SARAN**: Sequential Adaptive radial basis function network, which automatically optimize the center and scale of the basis function. Sparsity is encourage thorough $\mathcal{L}_1$ penalty. Both numerical studies and real-world

dataset from solar irradiance measurements demonstrate the superior performance of our method.

Another focus of this paper is to build accurate prediction of solar irradiance based on location (longitude and latitude) in the United States. We set up a statistical framework by using the proposed **SARAN** method to analyze a suite of data combining geographic measurements and two computer model outputs of solar irradiance at 1535 spatial locations across the United States. We showed that by ensemble predictions from measured data along with predictions from computer simulations, i.e. NAM and SREF, we can get a better prediction compared to just using measured data itself. Although forecast based computer model alone is not satisfying, we can fit surrogate model to extract large-scale spatial variability from the computer models to improve upon the accuracy obtained via fitting to the measured data alone. The final result can be used for a variety of real-world applications, e.g., identifying promising locations for solar farms.

Motivated by the encouraging results from this work, there are several interesting avenues for future work. One such direction is to model the solar irradiance based on both location and time. To do that we shall improve the scalability of **SARAN** to take advantage of the parallel computing framework. Another important aspect is to provide accurate predictive coverage for the **SARAN** method.

CHAPTER 4

ESTIMATED CAUSAL EFFECT OF PHYSICAL ACTIVITY PATTERN ON HEALTH OUTCOMES: A NONPARAMETRIC G-FORMULA APPROACH.

4.1 Abstract

4.1.1 Background

There is strong evidence that regular physical activity prevents premature death and several chronic diseases. In this paper, we assess the causal effect of time-varying physical activity pattern on the health outcomes BMI, vO2Max, SBP and quality of life.

4.1.2 Methods

Data from TRIPPA (trial of economic incentives to promote physical activity) was used to inform causal inference from physical activity groups to health outcomes. Principal components analysis and K-means clustering were performed to group the patients into clusters based on their activity patterns. We extended the traditional linear g-computation formula to a nonparametric g-computation formula to take into account non-linear time-varying confounding, and to estimate the causal effect of physical activity on health outcomes.

4.1.3 Results

After dimension reduction via principal components, clustering via K-means revealed four physical activity pattern groups, “active”, “inactive”, “weekday active”, and “weekend

active". Health outcomes were compared between "active" vs. "inactive" groups and "weekday active" vs. "weekend active" groups. Between the "active" and "inactive" groups, vO2Max was on average 1.43 higher in the "active" group compared to the "inactive" group (0.73, 2.21, $p < 0.001$), while quality of life score was on average 0.016 higher score in the "active" group compared to the "inactive" group (95% CI -0.001, 0.035; $p = 0.065$). There was no evidence suggesting differences in vO2Max or quality of life between the "weekday active" and "weekend active" groups. For BMI and SBP, there was no evidence suggesting differences between the four physical activity groups.

4.1.4 Conclusions

The results suggest a positive causal impact of membership in the "active" vs. "inactive" physical activity pattern group in terms of vO2Max.

4.2 Introduction

Non-communicable diseases are fast emerging as the major health challenge for the 21st century. Non-communicable diseases are responsible for nearly two-thirds of global death annually. [53] Physical inactivity is an important risk factor for non-communicable diseases and has been identified as the fourth leading risk factor for global mortality, contributing to 9% of deaths annually. [54] Regular physical activity can produce long-term health benefits. It can help prevent heart disease, cancer, improve heart and lungs condition. However, to the best of our knowledge, there is limited amount of research about the direct causal relationship between physical activity level and health outcomes. This dissertation chapter aims to study this important topic as a follow up study of TRIPPA (trial of economic

incentives to promote physical activity).

TRIPPA (trial of economic incentives to promote physical activity) was a four-arm, 6 month randomized controlled trial with a 6-month post-intervention follow-up period, conducted in 13 organizations spanning industries and sectors of government, to investigate the effects of an activity tracker, with or without cash or charitable incentives, on physical activity and health outcomes among full-time workers in Singapore. 800 participants were randomly assigned (1:1:1:1) with a computer generated assignment schedule to control (no tracker or incentives), Fitbit Zip activity tracker, tracker plus charity incentives, or tracker plus cash incentives (four arms), and had physical activity and health outcomes measured at baseline, 6 months and 12 months. Physical activity measures included daily steps and moderate-to-vigorous physical activity (MVPA) bout minutes per day. Health outcomes included systolic blood pressure (SBP), BMI, vO2Max and quality-of-life (QoL). Other covariates included gender, age, and ethnicity. [53] [55]

The goal of the original TRIPPA (trial of economic incentives to promote physical activity) study is to test the activity tracker, with or without incentives paid either in cash or via charitable donations, can increase physical activity and improve health outcomes among working people during a 6 month period, and to quantify which improvements are sustained during the subsequent 6 months after incentives are removed. Our study uses the same data to quantify the physical activity level and studies the causal relationship between physical activity level on health outcomes such as systolic blood pressure (SBP), BMI, vO2Max and quality-of-life (QoL). We hypothesized that higher the physical activity level (more steps) will result in better health outcomes. But due to the time frame of the TRIPPA study is only lasted for a year, the result may not be statistically significant.

4.3 Approach

Here, we assessed the causal effects of physical activity *clusters* on health outcomes. We first give a general overview of our approach then provide more detail for each step. Missing data was handled via multiple imputation. Specifically, 20 imputed datasets were generated. Physical activity patterns (daily steps and MVPA bout minutes) were dimension reduced via principal components[56] then clustered via K-means.[57] Finally, a non-linear g-formula approach is implemented to estimate the causal effect of physical activity cluster on health outcomes including SBP, BMI, vO2Max and QoL. Here, we treat each of the health outcomes (BMI, SBP, vO2Max and QoL) as the outcome of interest in turn, with the other health outcomes included as time-varying covariates.

Quantitative variables are summarized as mean (95% confidence interval) and categorical variables are summarized as rate (95% confidence interval). Means or rates are compared across clusters via F-tests or Fishers exact tests, as appropriate.

4.3.1 Multiple imputation

Multiple imputation [58] is a well-established and high-quality approach to handling missing data. It has been shown that multiple imputation is valid under missingness at random. The essential idea of multiple imputation is to generate several complete datasets with the missing data generated based on the estimated conditional distribution of the missing data given the observed data. The analyses are then conducted on each complete dataset and results are pooled across each of the complete datasets.

4.3.2 Physical activity clusters

Physical activity clusters were defined based on participants daily physical activity. Specific physical activity measures were daily steps and MVPA bout minutes over one week at baseline, six-month, and twelve-month follow-up, for 2400 total observations on the 14-dimensional weekly physical activity measure (steps and MVPA bout minutes each measured on each day of the week). Notably, physical activity cluster was examined from a time-varying perspective, where participants could change their physical activity cluster from baseline to 6 and 12-month follow-up. Principal components was used to reduce the dimension of the physical activity measures. The number of principal components was chosen to explain 90% of the total variance. Then, physical activity clusters were formed based on K-means clustering of their principal components. The number of clusters was chosen based on the elbow method with respect to total within-cluster sum of squares.

4.3.3 Causal inference

In attempting to measure the causal effect of time-varying physical activity cluster on health outcomes, it is important to consider the role of confounding variables that influence both the explanatory variable and the outcome. Much has been written on the general subject of confounding [59]. In this study, factors like BMI and weight potentially confound the causal relationship between the time-varying explanatory variable (physical activity cluster) and health outcomes like vO2Max. Further, these confounders themselves evolve over time and are measured repeatedly through the study. In this case, where time-varying confounders themselves affect the time-varying explanatory variable of interest, standard methods (like

propensity score matching) for dealing with confounding are difficult to apply[60].

We illustrate our scenario using a causal diagram [61] .The arrows in Figure 4.1 represent the assumed direction of causal influence. $A_0 - A_2$ represent the explanatory variables of interest, in our case the physical activity cluster at baseline, 6 months, and 12 months. We split the covariates into two groups, the first group represent the time-varying covariates including the BMI, SBP, and vO2Max and we denote them by $L_1 - L_3$. The second group represents the covariates that do not change over time such as gender, age, randomly assigned treatment arm, and ethnicity, and these are denoted by C. In this diagram, $Y_1 - Y_3$ are the outcomes of interest (QoL, for example) measured at each time point, and U denotes a set of unmeasured factors that influence both $L_0 - L_T$ and $Y_0 - Y_t$. Here, we assume that our scenario follows the no unmeasured confounders assumption, which states that conditional on $L_0 - L_T$ and $A_0 - A_T$, in the absence of causal effect of A_t on the Ys, A_t is independent of the Ys. U has arrows to all the variables except $A_0 - A_T$ and C has arrows to all the variables. These arrows are not included in the diagram for readability.

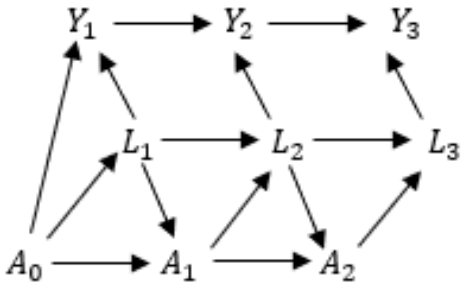

Figure 4.1: *Causal diagram for assessing the causal effect of physical activity cluster on health outcomes. As denotes time-varying causal variable of interest (physical activity cluster), Ls denotes time-varying confounders, and Ys denotes response of interest (health outcome). Time invariant confounders and shared influences on confounders and responses omitted for readability.*

A typical method for adjusting for confounding due to L is to condition on $L_0 - L_T$ in a

regression analysis. Unfortunately, that approach will not work in this situation. To identify the causal relationship between Y and A, a sufficient condition is the back-door criterion [62]. In brief, it says that if we want to assess the effect of A on Y and have a set of variables L as the control, then L satisfies the back-door criterion if (1) L blocks every path from A to Y that has an arrow into A, and (2) no node in L is a descendant of A. Suppose that we adjusted for $L_0 - L_T$, and we were interested in the causal effect of A_1 on Y_3. Controlling for L_1 has blocked the back-door path $A_1 \leftarrow L_1 \rightarrow L_1 \rightarrow Y_1 \rightarrow Y_2 \rightarrow Y_3$ However, controlling for L_2 has blocked the causal pathway from A_1 to Y_3. ($A_1 \rightarrow L_2 \rightarrow L_3 \rightarrow Y_3$).

Robins and colleagues have introduced several methods for estimating causal effects in the presence of time-dependent confounding. In this paper, we use a g-computation procedure [63] to estimate the causal effect of physical activity cluster on health outcomes. The g-formula works by first modeling the relationships between the variables in the observed data. Using these models, we simulate what would have happened to the subjects in the study had the variables $A_0 - A_t$ been fixed at specified levels, rather than been allowed to evolve naturally. The modeling and simulation are carried out forward in time. The process starts by modelling the time 1 data given the time 0 data, which allows simulation of the data at time 1 under various hypothetical interventions. Then, time 2 data are modeled given the time 0 and time 1 data in order to simulate the data at time 2 under various interventions, and so on. All post-baseline confounders and outcomes are simulated under specified levels of the causal variable of interest. Causal inference can then be pursued by comparing the outcomes under different specified levels as if these had been generated from a randomized experiment. The steps are as follows.

1. Developing models:

   (a) Develop a model for the conditional distribution of $L_1|A_0, C$

   (b) For each $t \in [2, T]$, develop a model for the conditional distribution of L_t given $L_0, A_0, , L_{t-1}, A_{t-1}, C$.

   (c) For each $t \in [2, T]$, develop a model for the conditional distribution of Y_t given $A_0, L_1, Y_1, , L_{t-1}, A_{t-1}, Y_{t-1}, C$.

2. Simulating under hypothetical settings:

   (a) Simulate L_t^* and Y_t^* incrementally from the conditional distributions developed above with $A_0, , A_T$ fixed at specified settings.

3. Comparing outcomes under hypothetical settings

   (a) Repeat the above simulations for a comparator $A_0, , A_T$ setting, and compare the distribution of outcomes under the two configurations.

4.3.4 Non-linear g-formula

In current literature, the vast majority of work using the g-formula approach has been restricted to linear parametric models due to the complexity of the final estimation [64] [65]. In particular, the g-formula estimation for the mean outcome at 12-month follow-up for a particular physical activity cluster setting is given by 4.1

$$E_c E_{L1|A_0,C} E_{Y_1|L_1,A_0,C...} E(Y_3|L_3, Y_2, L_2, Y_1, L_1, A_2, A_1, A_0, C) \tag{4.1}$$

If each of the component models is linear, the above estimation simplifies dramatically, while non-linear models necessitate modeling of the conditional distributions (above and beyond modeling the conditional means). In situations where we anticipate the possibility of non-linearities and complex interactions, machine learning techniques, such as random forest[66], can provide more accurate models for the component conditional distributions. Here, we develop a new framework using a modern predictive modeling perspective to estimate the g-formula estimation. The framework is similar to traditional g-formula but with some non-trivial modifications to estimate the full conditional distribution instead of only the mean. In particular, quantile regression forest [67] [68] is utilized to estimate the relevant conditional distributions in equation 4.1 above. Then simulated outcomes' are generated from the conditional distributions by uniformly randomly selecting the tenth (among the deciles), then generating a uniform draw on the selected tenth. The complete steps for performing non-linear g-formula are in the Appendix.

4.4 Results

After performing principle component analysis on the physical activity clusters, the first 9 PCs were selected to accurately represent our data (left panel of Figure 4.2). K-means clustering was then performed on the 9 selected PCs. The plot of within cluster total sums of squares vs. number of clusters suggested diminishing improvements in model fit beyond 4 clusters (right panel of Figure 4.2). Summary statistics for each of the identified 4 clusters are provided in Table 4.1.

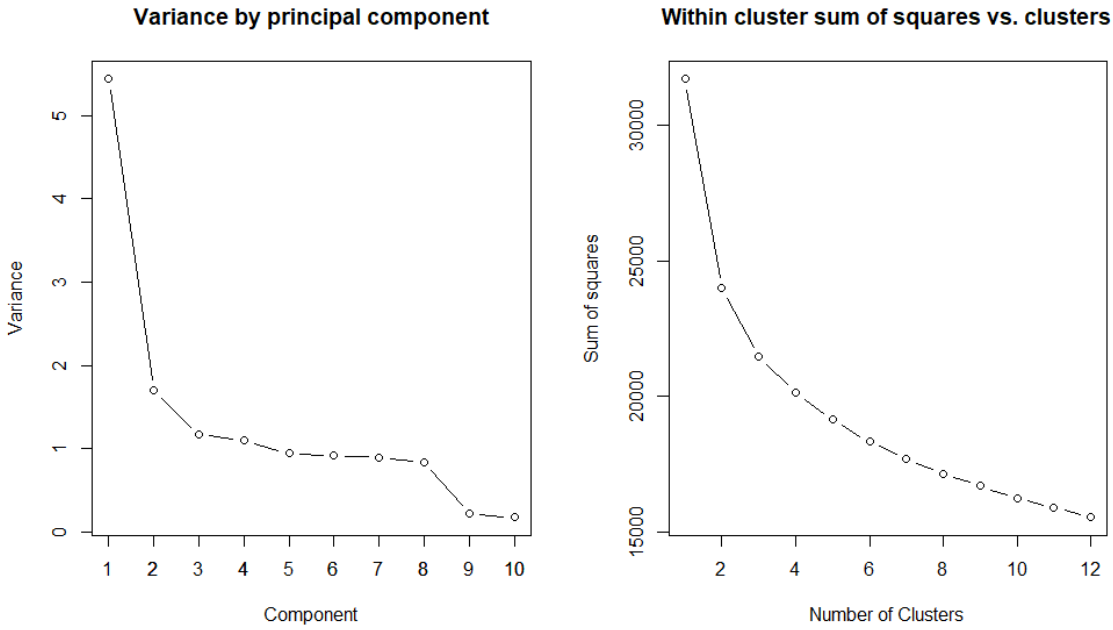

Figure 4.2: *(Left panel) principal component variance vs. number of principle components; (Right panel) total within cluster sum of squared errors vs. number of clusters.*

The clustering results suggest (Figure 4.3) that there is an active group, whose activity level is consistently relatively high and an inactive group whose activity level is consistently relatively low. The data also suggests that there are "weekend active" and "weekday active" clusters, who exercise more during the weekends and weekdays, respectively.

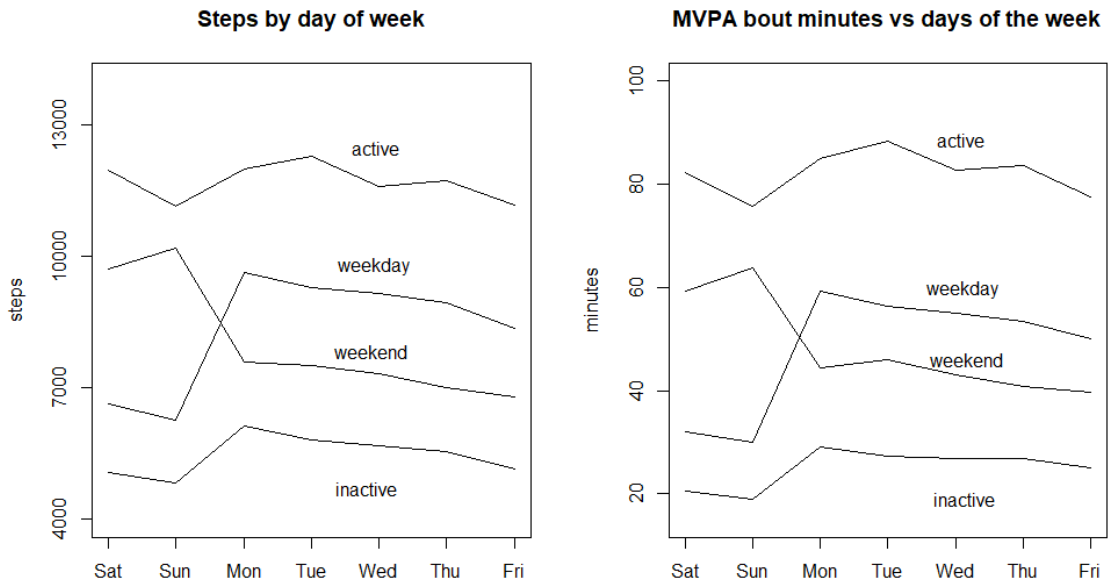

Figure 4.3: *Mean daily steps (left panel) and mean daily MVPA bout minutes (right panel) by physical activity level cluster.*

The baseline distributions of covariates and health outcome are summarized both overall and within physical activity clusters in Table 4.1. There is some evidence suggesting that the distributions of gender and baseline vO2Max differ between the four physical activity groups.

The causal inference results are summarized in Table 4.2. Mean differences are compared between the active group vs. "inactive" clusters and "weekday" vs. "weekend" clusters with respect to vO2Max, SBP, BMI and QoL. We find statistically significant evidence that membership in the "active" cluster has a positive causal effect on mean vO2Max as compared to the "inactive" cluster. There was suggestive evidence that membership in the "active" group had a positive causal impact on QoL as compared to the "inactive" group. There were no statistically significant causal effects between the "weekend" vs. "weekday" groups.

	Overall	Active	Weekdays	Weekend	Inactive	p-value
Age at baseline	35.5 (34.9,36.1)	37.0 (35.3,38.8)	34.4 (33.4,35.5)	35.3 (34.1,36.4)	36.0 (35.0,37.1)	0.784
Male gender	46.3% (42.8,50.0)	63.4% (54.3,72.5)	45.0% (38.8,51.2)	49.2% (41.9,56.5)	37.8% (31.8,43.8)	¡0.001
Chinese ethnicity	70.0% (66.8,73.1)	69.7% (59.1,76.6)	73.9% (68.4,79.4)	64.3% (57.4,71.3)	71.2% (66.7,77.9)	0.167
Weight at baseline (kg)	66.1 (65.1,67.1)	69.0 (66.3,71.7)	66.1 (64.3,67.9)	65.0 (63.1,67.0)	65.5 (63.8,67.3)	0.061
BMI at baseline	24.1 (23.8,24.4)	24.6 (23.8,25.4)	23.9 (23.4,24.5)	23.9 (23.3,24.4)	24.3 (23.8,24.8)	0.916
SBP at baseline (mmHg)	114.8 (113.7,116.0)	118.4 (115.3,121.7)	113.9 (111.8,115.9)	114.5 (112.5,116.6)	114.3 (112.1,116.4)	0.132
vO2Max at baseline	35.4 (35.1,35.7)	37.4 (36.6,38.2)	35.8 (35.5,36.4)	36.1 (35.5,36.7)	33.6 (33.1,34.2)	¡0.001
QoL at baseline	0.889 (0.879,0.897)	0.908 (0.884,0.931)	0.879 (0.862,0.895)	0.896 (0.877,0.914)	0.884 (0.868,0.900)	0.436

Table 4.1: *Participants characteristics at baseline overall and by physical activity group.*

	Mean difference	Confidence Interval	p-value
QoL (active vs inactive)	0.016	(-0.001,0.035)	0.065
QoL (Week vs Weekends)	0.002	(-0.008,0.012)	0.643
vO2Max (active vs inactive)	1.43	(0.73,2.21)	¡0.001
vO2Max (Week vs Weekends)	0.08	(-0.38,0.54)	0.752
BMI (active vs inactive)	-0.089	(-0.760,0.582)	0.801
BMI (Week vs Weekends)	-0.235	(-0.703,0.232)	0.324
SBP (active vs inactive)	-0.08	(-2.91,2.76)	0.960
SBP (Week vs Weekends)	-0.17	(-1.65,1.31)	0.824

Table 4.2: *Estimated causal effects (95% confidence intervals) for mean outcomes of interest for selected groups..*

The causal inference results suggest that membership in the active group has a positive impact on mean vO2Max compared to membership in the inactive group ($p < 0.001$) and suggests that the same relationship may also hold for quality of life ($p = 0.065$). This is as expected since people in the active group tend to exercise more and log more steps compared to the inactive group. We also find some evidence suggesting that people who exercise more tend may have better quality of life.

Although confounders are controlled in our analysis, it should be mentioned that our analysis is based on the assumption of no unmeasured confounders. Further, we also assumed that data was missing at random and multiple imputation results are valid.

4.5 Discussion

Physical exercises are known to be one of the best ways for people to stay healthy. However, our results find no strong evidence that higher physical activity level can lead to better health outcomes. Obviously, this study has several limitations. Because the candidates are volunteered to participate, the participants of TRIPPA study were probably healthier and more motivated to be physically active than the average full-time worker, so the selected samples may not be able to represent average people in the world. Another issue is the length of the TRIPPA study. Intuitively, increased level of physical exercises will not likely to have a significant impact on average persons health in less than a year. The longer-term benefit from physical exercise is more likely to be statistical significant than short-term.

In summary, we extended the original g-formula framework to include non-linear models, which allows us to use statistical models that are more robust compared to linear models. The follow-up study of TRIPPA is used to demonstrate the non-linear g-formula framework. The

results provide some evidence that a person who consistently spends more time exercising would tend to have higher vO2Max, and may have better quality of life, than if that person were to consistently remain relatively inactive. However, we find no evidence of causal differences in health outcomes for people who exercise more during the weekdays versus people who exercise more during the weekends.

Appendices

APPENDIX A

DETAILED NON-LINEAR G-FORMULA

Non-linear g-formula

1. Step 1:

   (a) Using quantile random forest to estimate quantile of conditional distribution $L_1|A_0, C$;

   (b) Using quantile random forest to estimate quantile of conditional distribution $Y_1|L_1, A_0, C$;

   (c) Using quantile random forest to estimate quantile of conditional distribution $L_2|Y_1, L_1, A_1, A_0, C$;

   (d) Using quantile random forest to estimate quantile of conditional distribution $Y_2|L_2, Y_1, L_1, A_1, A_0, C$;

   (e) Using quantile random forest to estimate quantile of conditional distribution $L_3|Y_2, L_2, Y_1, L_1, A_2, A_1, A_0, C$;

   (f) Using random forest to build regression model for $Y_3|L_3, Y_2, L_2, Y_1, L_1, A_2, A_1, A_0, C$;

2. Step 2:

   (a) Randomly select C from the observed data by bootstrapping;

   (b) Given the selected C and the A_{0k} of interest, randomly generate draw from

$L_1|C, A_{0k}$ by uniformly randomly select the tenth, then generate uniform draw on the selected tenth;

(c) Using similar procedures to generate Y_1, L_2, Y_2, L_3 by quantile distribution from step 1;

(d) Given the generated L_1, L_2, L_3, Y_1, Y_2, and A_{0k}, A_{1k}, A_{2k} of interest, compute $E(Y_3|L_1, L_2, L_3, Y_1, Y_2, A_{0k}, A_{1k}, A_{2k})$.

3. Step 3:

(a) Generate 500 bootstrapping samples for each multiple imputed dataset. Approximate $E_c E_{L_1|A_0,C} E_{Y_1|L_1,A_0,C...} E(Y_3|L_3, Y_2, L_2, Y_1, L_1, A_2, A_1, A_0, C)$;

(b) Using the multiple imputation pooling procedure to calculate if there is a statistically significant difference between different physical activity groups in terms of their health outcomes.

REFERENCES

[1] T. Ash, M. Ginty, and R. Page, *Landing Page Optimization: The Definitive Guide to Testing and Tuning for Conversions*, 2nd. Alameda, CA, USA: SYBEX Inc., 2012, ISBN: 0470610123, 9780470610121.

[2] G. Burtini, J. Loeppky, and R. Lawrence, "A Survey of Online Experiment Design with the Stochastic Multi-Armed Bandit", *arXiv e-prints*, arXiv:1510.00757, arXiv:1510.00757, 2015. arXiv: 1510.00757 [stat.ML].

[3] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules", *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, Mar. 1985.

[4] M. N. Katehakis and H Robbins, "Sequential choice from several populations", *Proceedings of the National Academy of Sciences*, vol. 92, no. 19, pp. 8584–8585, 1995. eprint: https://www.pnas.org/content/92/19/8584.full.pdf.

[5] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples", *Biometrika*, vol. 25, no. 3-4, pp. 285–294, Dec. 1933. eprint: http://oup.prod.sis.lan/biomet/article-pdf/25/3-4/285/513725/25-3-4-285.pdf.

[6] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, and Z. Wen, "A tutorial on Thompson Sampling", *Found. Trends Mach. Learn.*, vol. 11, no. 1, pp. 1–96, 2018.

[7] J. Langford and T. Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information", in *Advances in Neural Information Processing Systems*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds., Curran Associates, Inc., 2008, pp. 817–824.

[8] P. Auer, "Using confidence bounds for exploitation-exploration trade-offs", *Journal of Machine Learning Research*, vol. 3, no. 3, pp. 397–422, 2003.

[9] A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian Data Analysis*, 3rd. Chapman and Hall/CRC, 2013.

[10] S. Agrawal and N. Goyal, "Analysis of Thompson Sampling for the multi-armed bandit problem", in *Proceedings of the 25th Annual Conference on Learning Theory*, S. Mannor, N. Srebro, and R. C. Williamson, Eds., ser. Proceedings of Machine Learning Research, vol. 23, Edinburgh, Scotland: PMLR, 2012, pp. 39.1–39.26.

[11] A. J. C. Gittins and J. C. Gittins, “Bandit processes and dynamic allocation indices”, *Journal of the Royal Statistical Society, Series B*, vol. 41, no. 2, pp. 148–177, 1979.

[12] R. S. Sutton and A. G. Barto, “Reinforcement learning: An introduction”, *IEEE Transactions on Neural Networks*, vol. 16, pp. 285–286, 1998.

[13] R. Agrawal, “Sample mean based index policies with o(log n) regret for the multi-armed bandit problem”, *Advances in Applied Probability*, vol. 27, no. 4, pp. 1054–1078, 1995.

[14] S. L. Scott, “A modern Bayesian look at the multi-armed bandit”, *Appl. Stoch. Model. Bus. Ind.*, vol. 26, no. 6, pp. 639–658, Nov. 2010.

[15] S. Mak and C. F. J. Wu, “Analysis-of-marginal-Tail-Means (ATM): a robust method for discrete black-box optimization”, *Technometrics*, 2017, accepted.

[16] C. F. J. Wu, S. S. Mao, and F. S. Ma, “An investigation of OA-based methods for parameter design optimization”, Center for Quality and Productivity Improvement, University of Wisconsin-Madison, Tech. Rep. No. 24, 1987.

[17] ——, “SEL: A search method based on orthogonal arrays”, in *Statistical Design and Analysis of Industrial Experiments (S. Ghosh, ed.)*, Marcel Dekker, 1990, pp. 279–310.

[18] C. F. J. Wu and M. Hamada, *Experiments: Planning, Analysis, and Optimization*, 2nd. John Wiley & Sons., 2009, ISBN: 978-0-471-69946-0.

[19] V. R. Joseph, “A Bayesian approach to the design and analysis of fractionated experiments”, *Technometrics*, vol. 48, pp. 219–229, 2006.

[20] J. H. Albert and S. Chib, “Bayesian analysis of binary and polychotomous response data”, *Journal of the American Statistical Association*, vol. 88, no. 422, pp. 669–679, 1993.

[21] X. Li, N. Sudarsanam, and D. D. Frey, “Regularities in data from factorial experiments”, *Complexity*, vol. 11, no. 5, pp. 32–45, May 2006.

[22] S. P. Lloyd, “Least squares quantization in PCM”, *IEEE Transactions on Information Theory*, vol. 28, pp. 129–137, 1982.

[23] H. Su and C. F. J. Wu, “Funnel testing in webpage optimisation: Representation, design and analysis”, *Statistical Theory and Related Fields*, vol. 1, no. 1, pp. 3–14, 2017. eprint: `https://doi.org/10.1080/24754269.2017.1332964`.

[24] C. F. J. Wu, "Post-fisherian experimentation: From physical to virtual", *Journal of the American Statistical Association*, vol. 110, no. 510, pp. 612–620, 2015. eprint: https://doi.org/10.1080/01621459.2014.914441.

[25] D. J. FINNEY, "The fractional replication of factorial arrangements", *Annals of Eugenics*, vol. 12, no. 1, pp. 291–301, 1943. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1943.tb02333.x.

[26] H. Su and C. F. J. Wu, "Cme analysis: A new method for unraveling aliased effects in two-level fractional factorial experiments", *Journal of Quality Technology*, vol. 49, no. 1, pp. 1–10, 2017. eprint: https://doi.org/10.1080/00224065.2017.11918181.

[27] M. C. Kennedy and A. O'Hagan, "Bayesian calibration of computer models", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 3, pp. 425–464, 2001. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9868.00294.

[28] D. Higdon, M. Kennedy, J. Cavendish, J. Cafeo, and R. Ryne, "Combining field data and computer simulations for calibration and prediction", *SIAM Journal on Scientific Computing*, vol. 26, no. 2, pp. 448–466, 2004.

[29] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*. Springer-Verlag, New York, 2003.

[30] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, "Design and analysis of computer experiments", *Statistical Science*, vol. 4, no. 4, pp. 409–423, 1989.

[31] B. Haaland and P. Z. G. Qian, "Accurate emulators for large-scale computer experiments", *The Annals of Statistics*, vol. 39, no. 6, pp. 2974–3002, 2011.

[32] L. David and B. D.S., "Multivariable functional interpolation and adaptive networks", *Complex Systems*, vol. 2, no. 3, pp. 321–355, 1988.

[33] F. N. Melzi, T. Touati, A. Same, and L. Oukhellou, "Hourly solar irradiance forecasting based on machine learning models", in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2016, pp. 441–446.

[34] C. B. Martinez-Anido, B. Botor, A. R. Florita, C. Draxl, S. Lu, H. F. Hamann, and B.-M. Hodge, "The value of day-ahead solar power forecasting improvement", *Solar Energy*, vol. 129, pp. 192 –203, 2016.

[35] A. Alzahrani, J. Kimball, and C. Dagli, "Predicting solar irradiance using time series neural networks", *Procedia Computer Science*, vol. 36, pp. 623 –628, 2014, Complex Adaptive Systems Philadelphia, PA November 3-5, 2014.

[36] Z. Wang, F. Wang, and S. Su, "Solar irradiance short-term prediction model based on bp neural network", *Energy Procedia*, vol. 12, pp. 488 –494, 2011, The Proceedings of International Conference on Smart Grid and Clean Energy Technologies (ICSGCE 2011.

[37] J Zachariassen, K. Zeller, N. Nikolov, and T. McClelland, "A review of the forest service remote automated weather station (raws) network", *Technical report, General Technical Report RMRS-GTR-119*, pp. 1–161, Dec. 2003.

[38] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression", *Annals of Statistics*, vol. 32, pp. 407–499, 2004.

[39] R. Tibshirani, "Regression shrinkage and selection via the lasso", *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 58, pp. 267–288, 1994.

[40] C. T. R, *The r stats package*, R package version 3.5.2, 2018.

[41] T. Hastie, *Generalized additive models*, R package version 1.16, 2018.

[42] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'95, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1995, pp. 1137–1143, ISBN: 1-55860-363-8.

[43] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005, ISBN: 026218253X.

[44] G. M. Dancik and K. S. Dorman, "mlegp: Statistical analysis for computer models of biological systems using r", *Bioinformatics*, vol. 24, no. 17, pp. 1966–1967, 2008.

[45] R. B. Gramacy and D. W. Apley, "Local Gaussian process approximation for large computer experiments", *Journal of Computational and Graphical Statistics*, vol. 24, no. 2, pp. 561–578, 2015.

[46] R. B. Gramacy, "laGP: Large-scale spatial modeling via local approximate gaussian processes in R", *Journal of Statistical Software*, vol. 72, no. 1, pp. 1–46, 2016.

[47] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[48] W. S. Cleveland, E. Grosse, and W. M. Shyu, "Statistical models in s", in. Wadsworth & Brooks/Cole, 1992, ch. Local regression.

[49] H.-M. Kim, B. K. Mallick, and C. C. Holmes, "Analyzing nonstationary spatial data using piecewise gaussian processes", *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 653–668, 2005.

[50] C. J. Paciorek and M. J. Schervish, "Spatial modelling using a new class of nonstationary covariance functions", *Environmetrics*, vol. 17, no. 5, pp. 483–506, 2006. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/env.785.

[51] W. G. Cochran, "The combination of estimates from different experiments", *Biometrics*, vol. 10, no. 1, pp. 101–129, 1954.

[52] E. J. Green and W. E. Strawderman, "A james-stein type estimator for combining unbiased and possibly biased estimators", *Journal of the American Statistical Association*, vol. 86, no. 416, pp. 1001–1006, 1991.

[53] E. Finkelstein, B. Haaland, M. Bilger, A. Sahasranaman, R. Sloan, and E. Nang, "Effectiveness of activity trackers with and without incentives to increase physical activity (trippa): A randomised controlled trial", *The Lancet Diabetes & Endocrinology*, vol. 4, pp. 983–995, 12 Oct. 2016.

[54] I.-M. Lee, E. Shiroma, F. Lobelo, P. Puska, S. Blair, and P. Katzmarzyk, "Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy", *The Lancet*, vol. 380, pp. 983–995, 9838 Jul. 2012.

[55] E. Finkelstein, A. Sahasranaman, G. John, B. Haaland, M. Bilger, R. Sloan, E. Nang, and K. R Evenson, "Design and baseline characteristics of participants in the trial of economic incentives to promote physical activity (trippa): A randomized controlled trial of a six month pedometer program with financial incentives", *Contemporary clinical trials*, vol. 41, pp. 238–247, Feb. 2015.

[56] I. Jolliffe, *Principal Component Analysis*, 2nd. Springer Series in Statistics, 2009, ISBN: 978-0-387-95442-4.

[57] S. Lloyd, "Least squares quantization in pcm", *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[58] Y. C Yuan, "Multiple imputation for missing data: Concepts and new development", *SAS Institute Technical Report*, pp. 1–11, Jan. 2000.

[59] J. Pearl, "Causal inference in statistics: An overview", *Statistics Surveys*, vol. 3, pp. 96–146, Jan. 2009.

[60] R. Daniel, B. L. De Stavola, and S. N. Cousens, "Gformula: Estimating causal effects in the presence of time-varying confounding or mediation using the g-computation

formula", *The Stata Journal: Promoting communications on statistics and Stata*, vol. 11, pp. 479–517, Jan. 2012.

[61] S. Greenland, J. Pearl, and J. Robins, "Causal diagrams for epidemiologic research", *Epidemiology (Cambridge, Mass.)*, vol. 10, pp. 37–48, Feb. 1999.

[62] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference - Foundations and Learning Algorithms*, ser. Adaptive Computation and Machine Learning Series. Cambridge, MA, USA: The MIT Press, 2017.

[63] J. Robins, "A new approach to causal inference in mortality studies with a sustained exposure periodapplication to control of the healthy worker survivor effect", *Mathematical Modelling*, vol. 7, no. 9, pp. 1393 –1512, 1986.

[64] J. Snowden, S. Rose, and K. M Mortimer, "Implementation of g-computation on a simulated data set: Demonstration of a causal inference technique", *American journal of epidemiology*, vol. 173, pp. 731–8, Mar. 2011.

[65] J. G. Young, L. E. Cain, J. M. Robins, E. J. OReilly, and M. A. Hernn, "Comparative effectiveness of dynamic treatment regimes: An application of the parametric g-formula", *Statistics in Biosciences*, vol. 3, no. 1, p. 119, 2011, Exported from https://app.dimensions.ai on 2019/02/22.

[66] L. Breiman, "Random forests", *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[67] T. Hastie, *Quantile regression forests*, R package version 1.3-7, 2017.

[68] N. Meinshausen, "Quantile regression forests", *J. Mach. Learn. Res.*, vol. 7, pp. 983–999, Dec. 2006.

VITA

Vita may be provided by doctoral students only. The length of the vita is preferably one page. It may include the place of birth and should be written in third person. This vita is similar to the author biography found on book jackets.